Assessing modularity of developmental enhancers in Drosophila melanogaster

Assessing modularity of developmental enhancers in *Drosophila melanogaster*

A dissertation presented

by

Tara Laine Martin

to

The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Systems Biology

Harvard University
Cambridge, Massachusetts

August 2014

Assessing modularity of developmental enhancers in *Drosophila melanogaster*

Abstract

Gene expression is critical for animal development as cells divide and differentiate into multiple cell types. Cell-type specific gene expression is controlled by enhancers, DNA sequences that can direct expression of a target gene from hundreds of kilobases away. Gene promoters contact at least two enhancers on average, and enhancers may also contact each other. A key question is therefore how enhancers operate in this complex regulatory DNA context.

It has long been assumed that enhancers act as independent modules based on their ability to drive gene expression when isolated in reporter constructs. To test assumptions of enhancer modularity, I probed interactions between two developmental enhancers from the *even-skipped* locus in *Drosophila melanogaster* blastoderm embryos. My results contradict the classic definition of enhancers; I found that the arrangement of enhancers relative to one another and the promoter influences levels of gene expression while not affecting its spatial pattern within the embryo. These results are described in Chapter 2.

However, these enhancers are modular in one aspect: when fused directly together, they still direct their distinct spatial expression patterns. In Chapter 3 I describe a collaboration with Md Abul Hassan Samee in Saurabh Sinha's group at the University of Illinois Urbana-Champaign to apply computational sequence-to-expression models to my data. We found that a mechanistic model describing interactions between transcription factors was unable to fit our data well; in contrast, a phenomenological model that finds active sequences fits the data much better. These results indicate that to predict gene expression from sequence we will need to learn how enhancer boundaries are defined.

In summary, I present evidence that the organization of enhancers within a locus impacts expression of the target gene. This finding overturns assumptions about enhancer modularity and emphasizes the importance of considering higher level interactions across a locus. Structural variation is common in natural populations, and our results highlight a novel way in which these sequence variants may alter gene expression. To realize the long-standing goal to predict gene expression directly from sequence we must investigate how enhancers interact within a complex locus.

Table of Contents

Acknowledgments

I have so many people to thank for helping me through my thesis in all it's stages. It has been a long, arduous, confusing, fun, and amazing journey.

To my parents, thank you for always encouraging me to learn and celebrating my successes. To my partner, Dan, you are just incredible. In too many ways to count—but thank you for putting up with me! To so many friends—Kevin, Kitri, Karen, Sara, Michael, Jess, Zajj, and many more—thank you for your support and reminding me to keep celebrating life even when experiments are hard.

And of course, the DePace lab is such a wonderful place. I can't imagine a better group of people to work with; I'm going to miss you all terribly. Thank you especially to Kelly and Zeba for being the first ones to welcome me into the lab and make me feel at home. Also to Clarissa, Ben and Max for pushing and pulling me through the writing process, quite possibly the most exhausting thing I've ever done. I've always suspected Meghan is a superhero, and now I know that writing fly food emails is her super power.

Angela gave me an intellectual home at a critical point in my career. She is brilliant, creative, and visionary. I am so grateful that she took me in and proud that I got to be her first graduate student. There were some rocky points in figuring out how everything works, but I learned so much about both people and science and will always consider her a role model.

I would like to thank the members of my dissertation advisory committee for guidance: Marc Kirschner, Mitzy Kuroda, and Mike Springer. And the Systems Biology Program (especially Samantha Reed!) for generally being excellent and striving to constantly improve.

Finally, thank you to my collaborators Hassan Samee and Saurabh Sinha.

Chapter 1: Introduction

Overview

Enhancers are *cis*-regulatory sequences that direct cell-type specific gene expression during animal development. Because of their critical role in development, changes in enhancer sequences are associated with both morphological evolution (Wittkopp et al. 2009; Frankel et al. 2011; Mallarino et al. 2011; Jones et al. 2012) and disease (Karczewski et al. 2013; Maurano et al. 2012). Massive efforts have been focused on annotating and functionally characterizing enhancers (ENCODE Project Consortium et al. 2012; modENCODE Consortium et al. 2010; Kvon et al. 2014; Arnold et al. 2013; Dickel et al. 2014), and variation in non-coding regulatory sequences is common (Mu et al. 2011; Mackay et al. 2012). Therefore, learning how enhancers function is critical both for understanding development and for predicting the consequences of variation in regulatory sequences. However, analyzing the link between enhancer sequence and function is complicated by the fact that, in endogenous loci, enhancers are embedded in a complex regulatory environment where multiple enhancers and other types of *cis*-regulatory sequences are present (Bulger and Groudine 2010).

In order to understand how enhancers function in their endogenous context, we need to better understand how enhancers interact with each other. In species as diverse as humans and fruit flies most genes are contacted by multiple enhancers (Jin et al. 2013; Ghavi-Helm et al. 2014). Surveys have identified at least 400,000 putative enhancers in the human genome (ENCODE Project Consortium et al. 2012) and measurements of chromatin contact frequency indicate that enhancers interact with each other as well as gene promoters (Jin et al. 2013). Since many enhancers are only detectable in particular cell types (Buecker and Wysocka 2012), these measurements are likely to prove to be underestimates as more cell types and

developmental stages are surveyed. However, the functional consequences of enhancer-enhancer interactions are not known.

Enhancers have traditionally been defined as independently acting modules (reviewed in Bulger and Groudine 2010; Ong and Corces 2011; Buecker and Wysocka 2012). This definition enables complex gene expression patterns to be studied by interrogating each enhancer individually. The regulatory sequences which control a gene can span thousands of kilobases and produce complex patterns of expression in different tissues at different developmental stages (Levine 2010). The existence of enhancers which direct only a subset of the total expression pattern enabled researchers to break the problem of deciphering regulatory sequences into more manageable chunks (Fujioka et al. 1999). Studies of single isolated enhancers have produced numerous insights into detailed transcription factor control of transcription (Arnosti et al. 1996; Li and Arnosti 2011; Swanson et al. 2010), leading to predictive models of enhancer function (Janssens et al. 2006; Segal et al. 2008; He et al. 2010).

However, if enhancers do not act independently, these studies of enhancers in isolation will not scale back up to predict gene expression at the level of complex endogenous loci. In fact, recent attempts to scale up computational models of single enhancer function to regulation of gene expression in an endogenous locus have encountered challenges (Kim et al. 2013; Samee and Sinha 2014). Predicting gene expression in an endogenous context is a key goal for understanding the effects of structural genetic variants which can change spacing, copy number and order of enhancers within a locus (Pang et al. 2010).

The experiments cited in support of the modular definition of enhancers demonstrate the remarkable flexibility of enhancers to drive expression from heterologous promoters and from widely varying genomic positions (Banerji et al. 1981). However, these experiments are not a direct test of enhancer modularity. Enhancer modularity requires three primary characteristics: they must have strict DNA sequence boundaries, enhancers must act independently of distance

and orientation, and the output of a locus must be the sum of the activity of multiple enhancers (Figure 1.1). In this Chapter, I will discuss the historical roots of modularity in each of these contexts, recent conflicting evidence, and how my thesis tested the modular definition of enhancers directly.

Enhancer boundaries

The first identified enhancers drove gene expression in an inducible fashion; they were short and had well defined boundaries. For example, the viral SV40 enhancer consists of two 72bp sequence repeats which synergistically bind a pair of activating TFs (Banerji et al. 1981; Levine 2010), and the virus inducible IFN-beta enhanceosome consists of a 55bp sequence containing eight tightly spaced binding sites for six TFs (Panne et al. 2007). The enhanceosome is defined by its cooperative binding of activators which creates switch-like gene expression in response to viral infection. The assembly of TF complexes on the DNA is integral to the function of these enhancers and precisely defines the boundaries. However, these functional constraints are not typical of other enhancers.

Compared to inducible enhancers, developmental enhancers are longer and have more flexible TF binding site arrangements. For example, in *Drosophila melanogaster* embryos, an enhancer which drives expression of *even-skipped (eve)*, a key developmental gene, displays high levels of TF binding site turnover between species while maintaining function (Ludwig et al. 2000; Hare et al. 2008). Flexible sequence arrangements such as seen in the *eve* stripe 2 enhancer are referred to as “billboard” enhancers (Arnosti and Kulkarni 2005), and they appear to be quite common (Arnold et al. 2014). The ability of billboard enhancers to function with different compositions and arrangements of TF binding sites suggests that the boundaries of these enhancers are less well defined. Attempts to shorten the enhancers to minimal sequences can result in weak or ectopic expression, but multiple sequences with minimal

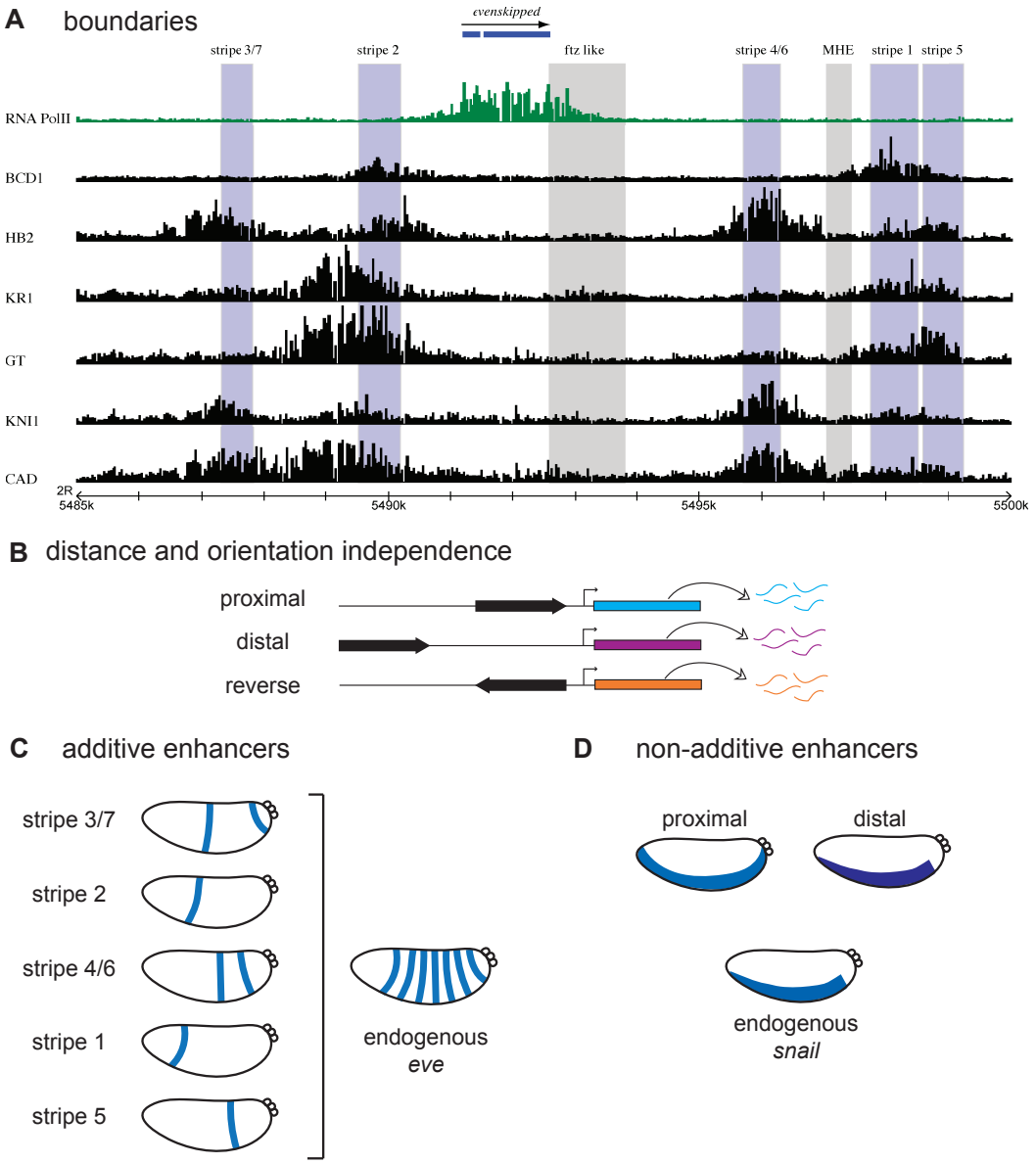


Figure 1.1: Modularity is associated with three primary characteristics of enhancers. A) Enhancer boundaries are sharper than TF occupancy profiles. The *eve* locus is illustrated with the stripe enhancers shaded in blue and enhancers active later in development shaded in tan. TF occupancy profiles are shown in black for known regulators of *eve* expression. Occupancy was measured using ChIP-chip, adapted from (Li et al. 2008). B) Enhancers are defined as acting independent of distance and orientation relative to the promoter. In this cartoon the thick black arrows represent an enhancer that has been placed upstream of the promoter driving a reporter gene (blue, purple or orange). C) The standard model for *eve* regulation in the blastoderm is that five stripe enhancers each drive one or two stripes of expression (blue). The endogenous pattern is a sum of the enhancer activities. D) The endogenous *snail* pattern is driven by two enhancers which interact in a non-additive fashion both in terms of level and position (Dunipace et al. 2011). The proximal enhancer drives expression in the tail, which is repressed in the presence of the distal enhancer. Meanwhile, the distal enhancer alone drives much higher expression than the endogenous gene, suggesting that the proximal enhancer limits overall levels through an unknown mechanism.

overlap are able to drive similar robust expression patterns (for example see comprehensive dissection of the *eve* locus in Fujioka et al. 1999). Reciprocally, flanking sequences can contribute to enhancer activity (such as robustness to environmental perturbation), even if they don't have activity on their own. In fact, binding sites in surrounding sequences can contribute to enhancer activity and robustness to environmental perturbation (Ludwig et al. 2011).

Genomic measurements of enhancer features, such as TF occupancy and DNA accessibility, also indicate that developmental enhancer boundaries may not be strict cut-offs (Figure 1.1A; adapted from Li et al. 2008). The spread in these measurements could be due to weakly defined enhancer boundaries or technical limitations that obscure sharp signals. The resolution along the DNA for genomic measurements is limited by DNA fragment size in the case of sequencing based methods, or length of probes for microarray based measurements. In addition, genomic measurements must average over many cells and time points in the embryo and over time when tissue is homogenized; this can lead to the appearance of soft boundaries even if they are sharp within single cells at precise times.

Independence of distance and orientation

The definition of enhancers as elements that act independently of distance and orientation with respect to the promoter is based on an indirect experiment: a candidate enhancer is inserted into a plasmid containing a basal promoter and reporter gene and expression is measured using either *in situ* hybridization in embryos or an enzymatic assay in cell culture. Critically, neither measurement accurately captures expression levels. The first characterized enhancer was the SV40 enhancer, which normally drives expression of viral genes but was shown to activate a beta-globin reporter from 10kb away (Banerji et al. 1981). This was followed shortly by the discovery of enhancers in the mammalian genome at the immunoglobulin locus (Banerji:1983wc; Gillies et al. 1983). The discovery that metazoan

regulatory sequences could act at a distance was revolutionary, and sparked the hunt for distal acting regulatory sequences that continues to this day (ENCODE Project Consortium et al. 2012; Kvon et al. 2014).

The definition of enhancers that emerged from these experiments overstates the case for modularity. These early studies were transient transfection assays in cell culture with limited quantitative resolution. Even so, distance dependent effects on reporter expression level were observable (Moreau:1981uw; Wasylyk et al. 1984). It would be more accurate to say that enhancers are capable of acting across different distances and orientations, though the precise gene expression levels and patterns they drive may change with context.

Interactions between enhancers

The ability of enhancers to act at a distance, combined with the dissection of developmental loci into fragments that each drove a different portion of total gene expression, led to a model of enhancers as modular components that act additively to produce the overall gene expression pattern. Modularity has been proposed to enable complex regulatory control and the reuse of genes in different cellular contexts without risk of pleiotropy (Kirschner and Gerhart 1998; Carroll 2000; Levine 2010). A gene regulated by multiple modular enhancers would be able to respond to different signaling inputs under different contexts without needing to integrate all these signals within the same function.

A canonical example of a modular developmental locus is the *Drosophila melanogaster* *eve* locus, a pair-rule gene (Figures 1.1C and 1.2). The *eve* gene is expressed in seven stripes along the AP-axis during embryogenesis under the control of five enhancers which each drive expression of one or two stripes (Small et al. 1991; 1996; Fujioka et al. 1999). Additional enhancers control expression at later developmental stages in a variety of tissues, including motor neurons (Fujioka et al. 1999). By reducing the complex seven stripe pattern of *eve* into

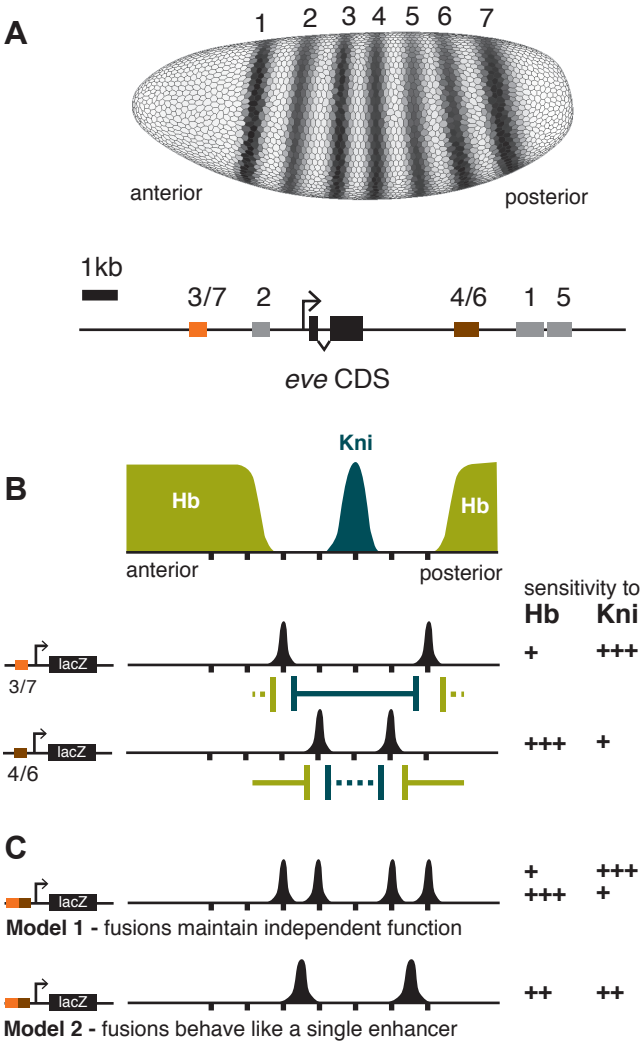


Figure 1.2: A) The *eve* locus contains five stripe enhancers encoding the seven stripe pattern of expression in blastoderm embryos. B) *Eve* is regulated by the repressors *hb* (green) and *kni* (blue). The boundaries of expression of eve 3/7 and eve 4/6 (black peaks) are set by differential sensitivities (dashed and solid lines, and table on right) to each repressor (cartoon adapted from Clyde *et al*, 2003). C) Information integration of the fusions at different length scales predicts different expression patterns driven by the fusions. If the component enhancers remain autonomous (model 1), a four stripe pattern is produced because two distinct sensitivities to *hb* and *kni* are maintained. If the entire fusion is interpreted as a single window (model 2), only two stripes are expressed as the result of a single sensitivity to *hb* and *kni*. Intermediate models are also possible but not shown.

component patterns, researchers were able to pick apart the detailed transcription factor binding responsible for regulating *eve* expression (Arnosti et al. 1996; Clyde et al. 2003; Struffi et al. 2011). Although transgenic assays were limited to identification of qualitative patterns, they were successful at dissecting many developmental loci into component parts for further study (Schroeder et al. 2004; Segal et al. 2008; Gallo et al. 2011). Building on discoveries made in individual loci, genomic approaches have gone on to identify characteristics of DNA accessibility, chromatin marks and TF occupancy that are associated with enhancers. These signals have in turn been used to identify enhancers genome wide—hundreds of thousands of

enhancers have now been identified in various human cell types, most of them cell type specific (reviewed in Buecker and Wysocka 2012).

A number of loci cannot be dissected into modular enhancers. For example, in the *sloppy-paired* (*slp1*) locus, two enhancers are required to drive the endogenous expression pattern, but they do not produce this pattern additively; instead one enhancer produces some of the *slp1* stripes while the distal early stripe element (DESE) expresses all stripes with additional ectopic expression in inter-stripe regions (Fujioka 2010). In other *Drosophila* developmental gene loci a number of "shadow" enhancers have been found which display overlapping expression patterns to previously identified enhancers (Perry et al. 2010). Several of these enhancers show non-additive interactions in terms of either spatial expression (*hunchback*, *knirps* (Perry et al. 2011), *snail* (Dunipace et al. 2011)) or level (*snail* (Dunipace et al. 2011)). Finally, measurements of chromatin conformation suggest that many enhancers interact directly with each other (Montavon et al. 2011; Li et al. 2012), which may provide a mechanism for non-additive behavior. The question that naturally arises is whether the modular behavior originally attributed to enhancers is the norm, or whether the growing list of exceptions represents a more realistic view of enhancer biology.

Testing enhancer modularity

I sought to directly challenge the definition of enhancers as "modules" by rearranging two well characterized enhancers relative to the promoter and one another. I chose enhancers from the *eve* locus, the canonical example of a modular developmental locus in *Drosophila melanogaster* embryos. The design of the experiments allowed me to test all three characteristics of enhancer modularity—boundaries, distance and orientation, and interactions with other enhancers. I quantitatively measured the effects of distance and orientation relative to the promoter on levels of expression driven by each of two enhancers using *in situ* hybridization

and high-resolution imaging. I also compared the expression levels driven by fusions of the two enhancers, with or without intervening spacer sequences, to simultaneously test the precision of enhancer boundaries and additivity of expression patterns.

The *Drosophila* blastoderm embryo is an ideal system for studying the mechanism of enhancer function: it is patterned by a well defined developmental gene network and quantitative measurements are technically tractable. Seminal genetic screens defined transcriptional cascades that pattern the anterior-posterior and dorsal-ventral axes (Nüsslein-Volhard and Wieschaus 1980). The TFs in these cascades have been extensively characterized; their DNA binding preferences are known (Zhu et al. 2011; Bergman et al. 2005) and their interactions with target genes have been largely defined (MacArthur:2009io; Li et al. 2008). Patterning occurs rapidly over the first 4 hours of development, and the embryo is amenable to quantitative imaging techniques that measure both the expression of regulating TFs and their targets (Luengo Hendriks et al. 2006; Fowlkes et al. 2008; Pisarev et al. 2009) as well as genetic manipulation of regulatory sequences (Venken et al. 2009). The powerful combination of extensive knowledge of the patterning network, quantitative data, and genetic manipulability provides a foundation for testing principles of gene regulation.

The *eve* enhancers provide an excellent model system for testing hypotheses about enhancer modularity. The eve 3/7 and eve 4/6 enhancers are extremely well characterized with known regulators verified through both *trans* and *cis* mutation experiments (Fujioka et al. 1999; Clyde et al. 2003; Struffi et al. 2011). Previous fusions of the eve 2 and eve 3/7 enhancers concluded that changes in expression level were due to local TF-TF interactions at the boundaries of enhancers (Small et al. 1993; Kim et al. 2013). However, these experiments tested only two arrangements of the two enhancers relative to one another and used endogenous flanking sequences as “spacers.” As can be seen in Figure 1.1A the sequences flanking the eve 2 and eve 3/7 enhancers exhibit high TF occupancy, and Kim et al. (2013)

concluded that TF binding sites within the flanking sequences functionally contributed to the expression patterns. I chose to use synthetic arrangements with "neutral" spacer sequences taken from a bacterial coding gene between the two enhancers to attempt to isolate the consequences of distance, orientation, and order of enhancers from possible TF-TF interactions. In addition, I tested four different fusions that cover all possible junctions between the two enhancers.

The eve 3/7 and eve 4/6 enhancers share two repressors, *hunchback* (*hb*) and *knirps* (*kni*), but they respond to them with different sensitivities to produce distinct pairs of stripes in stereotyped positions (Small et al. 1996; Fujioka et al. 1999; Clyde et al. 2003; Struffi et al. 2011)(Figure 1.2B). The fact that different sensitivities to these repressors results in different spatially resolved expression patterns is very useful for our experiments: we can qualitatively assess whether the fusion expression patterns result from a single sensitivity to these repressors by looking at the position of the resulting expression pattern. We can also quantitatively assess whether a single sensitivity to the shared regulating TFs can explain the expression pattern by modeling the relationship between regulating TFs and the expression pattern computationally.

My experiments unequivocally demonstrate that enhancers are not modular; these results are described in Chapter 2. First, the level of expression driven by enhancers changes with distance and orientation to the promoter. Second, despite changes in level, the position of expression driven by fused enhancers does not change as expected, indicating that enhancer boundaries are maintained in the fusions. Third, there are strong non-additive interaction effects between enhancers. In particular, the levels of expression driven by each enhancer change dramatically depending on their order relative to the promoter.

In Chapter 3, we further probe the boundaries of the two enhancers using computational statistical thermodynamic models to test how well known mechanisms of TF interactions can

explain the observed spatial expression patterns. We find that local interactions between TFs are insufficient to explain the observed modularity of fused enhancers.

I discuss the implications of this new view of enhancer modularity and future directions of study in Chapter 4.

Chapter 2: Locus architecture affects mRNA expression levels in *Drosophila* embryos

Tara Lydiard-Martin, Meghan Bragdon, Kelly B. Eckenrode, Zeba B Wunderlich, Angela DePace

Author Contributions

TLM and AHD designed the experiments. TLM performed all experiments and image analysis. MDB and KE assisted with fly husbandry, embryo collection, *in situ* hybridizations and imaging. ZBW processed raw image files into pointcloud files for further analysis. TLM and AHD wrote the text.

Abstract

Structural variation is common in the genome due to insertions, deletions, duplications and rearrangements. However, little is known about the ways structural variants impact gene expression. Developmental genes are controlled by multiple regulatory sequence elements scattered over thousands of bases; developmental loci are therefore a good model to test the functional impact of structural variation on gene expression. Here, we measured the effect of rearranging two developmental enhancers from the *even-skipped (eve)* locus in *Drosophila melanogaster* blastoderm embryos. We systematically varied orientation, order, and spacing of the enhancers in transgenic reporter constructs and measured expression quantitatively at single cell resolution in whole embryos to detect changes in both level and position of expression. We found that the position of expression was robust to changes in locus organization, but levels of expression were highly sensitive to the spacing between enhancers and order relative to the promoter. Our data demonstrate that changes in locus architecture can dramatically impact levels of gene expression. To quantitatively predict gene expression from sequence, we must consider how information is integrated both within enhancers and across the gene locus.

Introduction

How do changes in regulatory DNA sequence impact gene expression? This question is critical for understanding metazoan disease and evolution because precise control of gene expression is necessary for the development and function of metazoan cells. Mis-regulation is increasingly implicated in a broad range of disease states (Karczewski et al. 2013; Maurano et al. 2012), and changes in gene expression underlie some morphological differences between animal species (Wittkopp et al. 2009; Frankel et al. 2011; Mallarino et al. 2011; Manceau et al. 2011; Jones et al. 2012). Natural variation in regulatory DNA is common (Mu et al. 2011; Mackay et al. 2012), but not all changes in regulatory sequence have functional consequences (Romano and Wray 2003; Hare et al. 2008; Swanson et al. 2011). A central challenge is to learn which, and to what extent, regulatory sequence variants alter gene expression.

Many classes of *cis*-regulatory elements that direct metazoan gene expression have been identified, including enhancers, silencers, insulators and targeting sequences (Maston et al. 2006). Cell type specific expression is primarily directed by enhancers that integrate information from multiple DNA-bound transcription factors (TFs) to produce a specific expression pattern (reviewed in Bulger and Groudine 2010). These short (~1kb) sequences can be located upstream, downstream, or within introns of their target gene. Many genes, particularly key developmental transcription factors, are regulated by several enhancers that together direct the total gene expression pattern (Levine 2010; de Laat and Duboule 2013); accordingly, mutation or loss of enhancer sequences can have phenotypic consequences (VanderMeer and Ahituv 2011; Dunipace et al. 2011; Kim et al. 2014).

Natural variation in regulatory sequence spans multiple length scales, from single nucleotide polymorphisms (SNPs) to structural variants such as insertions, deletions, duplications, inversions, and translocations that can range in size from 1-10bp “micro-indels” up through 1Mb (Pang et al. 2010). In humans, structural variation is estimated to account for more

than 10 times as much genomic variation between individuals as SNPs (Pang et al. 2010). Specific examples of structural variants have been associated with disease (Kleinjan and Coutinho 2009) and morphological evolution (Jones et al. 2012). Structural variants appear to be under strong purifying selection pressure; structural variants in non-coding sequences are selected against more strongly than non-synonymous base substitutions in coding sequences (Zichner et al. 2013).

Despite the prevalence of structural variation, the consequences of large scale regulatory rearrangements for gene expression have not been systematically studied. Many studies of regulatory sequence variation have focused on the functional impact of SNPs and small indels, either by directed mutagenesis (Thanos and Maniatis 1995; Arnosti et al. 1996; Swanson et al. 2010), or systematic characterization of enhancer variant libraries (Erceg et al. 2014; Melnikov et al. 2012; Kwasnieski et al. 2012; Smith et al. 2013; White et al. 2013). These studies have elucidated how sequence changes within an enhancer impact its regulatory function. Structural variants, meanwhile, may influence the expression of a gene by changing the relative contributions from different enhancers without altering the individual enhancer functions. Most simply, deleting enhancers can disrupt gene expression (Ludwig et al. 2005; Guenther et al. 2008; Chan et al. 2010; Dunipace et al. 2011; Montavon et al. 2011). Conversely, enhancer duplications also impact gene expression, but in unpredictable ways (Klopocki et al. 2008). Rearrangements that move enhancers relative to one another may also alter expression if their bound TFs interact (Small et al. 1993; Kim et al. 2013). Finally, structural variants might disrupt the 3D structure of a locus, which changes during development (Kagey et al. 2010; Phillips-Cremins et al. 2013) and is important for the regulation of gene expression (Deng et al. 2012; Dekker et al. 2013).

To investigate how structural variants impact gene expression, we created a set of reporter constructs in which we systematically vary the orientation, order, and spacing between

two enhancers. TFs are known to interact through short-range and long-range repression mechanisms (Gray and Levine 1996; Courey and Jia 2001; Li and Arnosti 2011), we therefore tested a series of distances between enhancers spanning 0 to 1000bp. We chose to conduct this study in *Drosophila melanogaster* blastoderm embryos because we could 1) use two well-characterized enhancers from the highly studied *even-skipped* (*eve*) locus (Fujioka et al. 1999; Clyde et al. 2003; Struffi et al. 2011); 2) readily integrate our reporters *in vivo* (Groth et al. 2004); and 3) make quantitative measurements of expression at cellular resolution using fluorescent imaging (Fowlkes et al. 2008; Wunderlich et al. 2014). This powerful system allowed us to quantitatively probe enhancer activity in the full range of cell types present in developing embryos.

Our results demonstrate that structural variants can have a strong effect on gene expression level. First, contrary to the classic definition of enhancers, we found that levels of expression driven by single enhancers vary with orientation and distance to the promoter; the magnitude and direction of this effect was enhancer-specific. Second, in configurations containing two enhancers, expression pattern position was largely maintained but levels of expression varied by nearly 8-fold depending on the orientation, order and spacing of the enhancers relative to one another and the promoter. Third, we found that output driven by two enhancers is not equivalent to additive output from the two component enhancers, even when they are separated by a 1000bp neutral spacer sequence; this indicates enhancers can interact at a much longer range than previously reported. Taken together, our results suggest that structural variants that alter locus architecture are likely to have a substantial impact on gene expression levels. These results emphasize that, in order to quantitatively predict gene expression from sequence, we must consider how information is integrated at multiple scales—both within enhancers and across gene loci.

Results

We chose two well-characterized enhancers from the *eve* locus for our study: eve 3/7 (which drives expression of stripes 3 and 7) and eve 4/6 (which drives expression of stripes 4 and 6). These two enhancers share regulators (Clyde et al. 2003) and are normally located on opposite sides of the locus (Figure 2.1A). We engineered various arrangements of these two enhancers to each other and the promoter, using typical reporter constructs that contain the *eve*

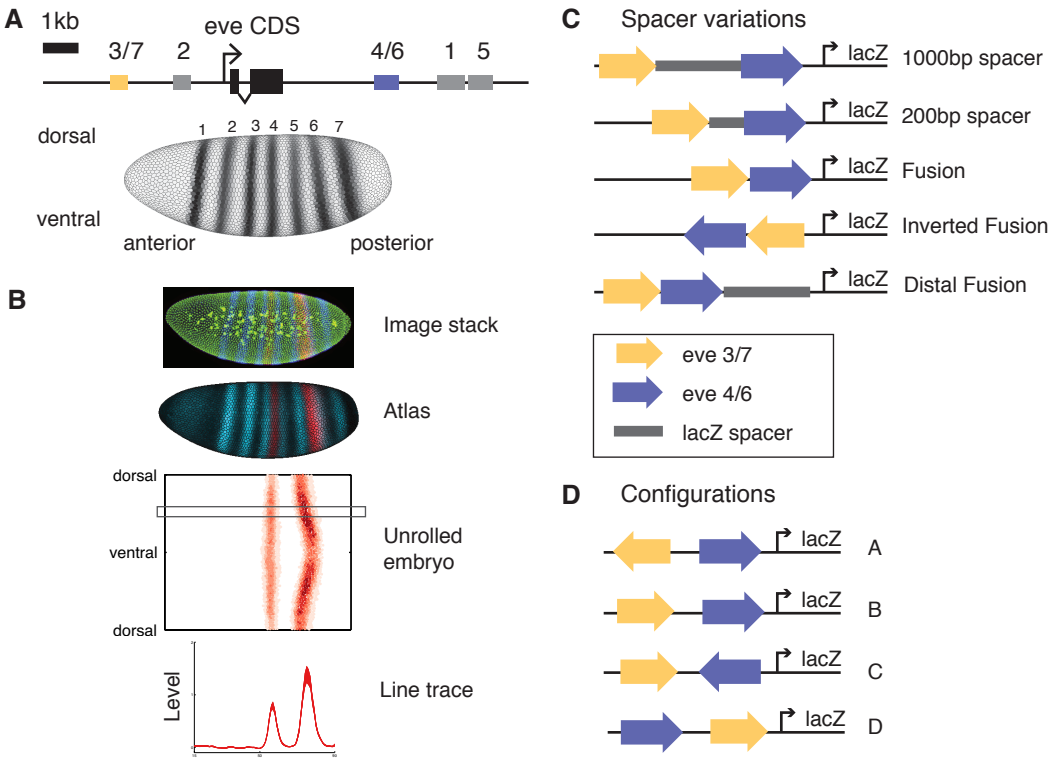


Figure 2.1: Synthetic arrangements of two enhancers from the *eve* locus test consequences of genomic structural variants A) The *eve* locus contains five stripe enhancers encoding the seven stripe pattern of expression in blastoderm embryos. B) We stained embryos for a reporter gene (red) using fluorescent *in situ* hybridization, and collected image stacks through the entire embryo. We computationally segmented embryos and extracted fluorescence values for each cell, then aligned embryos to an average morphological framework to generate an atlas of average expression patterns (see Materials and Methods). During the hour of development under study the cells are in a sheet on the surface of the embryo and can be represented in 2D as an unrolled cylindrical projection. For simplicity, in most figures we show a subset of our data taken from a line trace through the lateral side of the embryo (grey box). C) We tested synthetic arrangements of two enhancers with different length and positioning of spacers. D) We also tested different configurations of the two enhancers that cover all possible junctions between the two.

basal promoter driving expression of lacZ (Hare et al. 2008). We integrated these reporters into the same genomic location using the phiC31 site-directed integration system (Groth et al. 2004; Fish et al. 2007). When spacer sequence was required, we used portions of the lacZ coding sequence chosen to minimize predicted binding sites for the regulators of these two enhancers (Supplemental Figure 2.1 and Supplemental Figure 2.2).

Enhancer distance and orientation relative to the promoter affect target gene expression quantitatively

We first measured the effect of changing a single enhancer's distance and orientation from the promoter. We cloned the minimal eve 3/7 (511bp; Small et al. 1996) and eve 4/6 (800bp; Fujioka et al. 1999) enhancers at three positions (0bp, 500bp, and 1000bp upstream of the promoter) and in two orientations (either the endogenous orientation, or reversed). We measured the expression from each reporter construct in blastoderm embryos using fluorescent *in situ* hybridization against the lacZ reporter gene and an endogenously expressed fiduciary marker, *fushi-tarazu* (*ftz*). We imaged entire embryos at cellular resolution and assembled our data into a gene expression atlas, which contains average levels of expression for each gene in each cell for six time points during the hour prior to gastrulation (Fowlkes et al. 2008). To normalize levels of expression across reporter constructs we co-stained reporter lines with the endogenous gene *huckebein* (*hkb*) in the same channel as lacZ (Wunderlich et al. 2014). For simplicity, in most figures we show a lateral line trace—the moving average of expression level for a five nuclei wide dorsal-ventral (D/V) strip along the anterior-posterior (A/P) axis—for the third time point.

We anticipated that these constructs would merely serve as controls for more complex rearrangements of two enhancers relative to one another, but we found that rearrangements of single enhancers have significant effects on the level, but not the position, of expression.

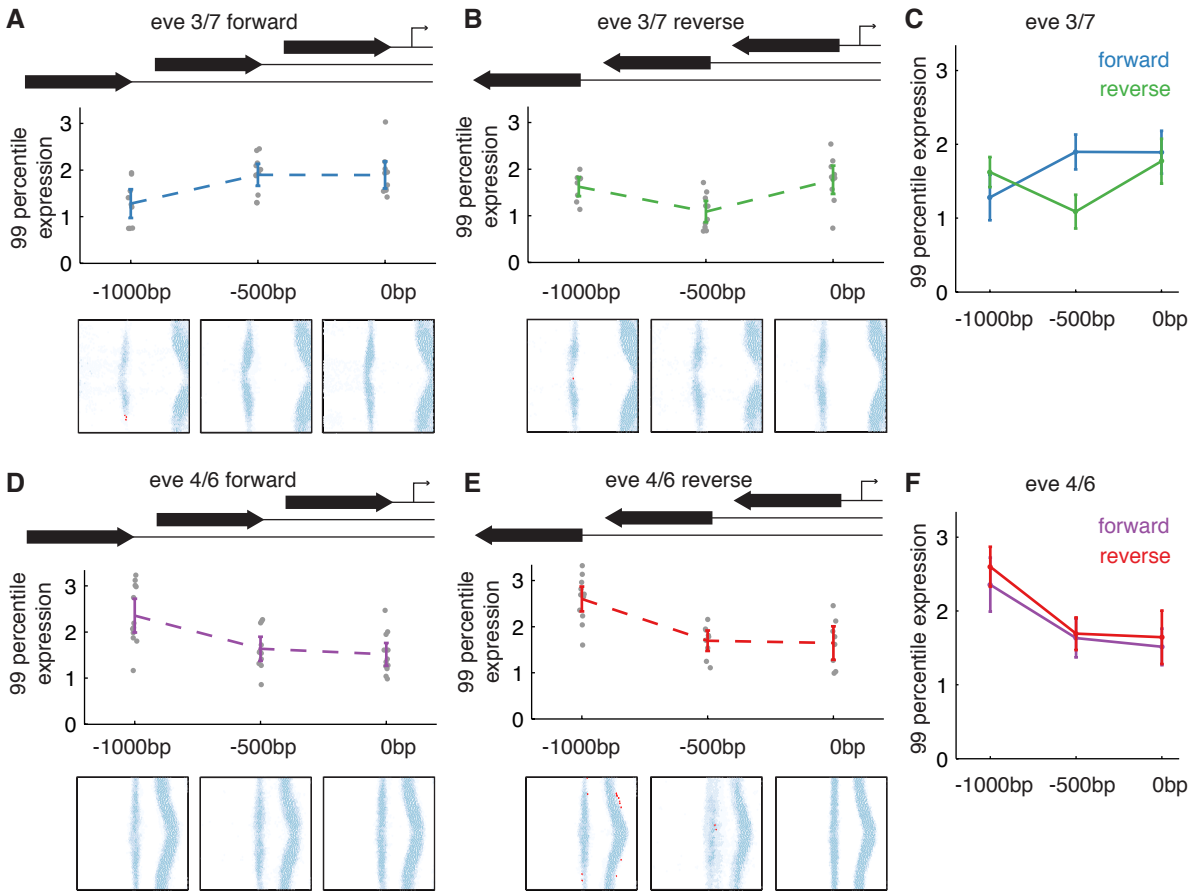


Figure 2.2: Expression levels driven by the eve 3/7 and 4/6 enhancers depend on enhancer position and orientation relative to the promoter. We measured expression driven by the eve 3/7 (A, B) and eve 4/6 (D, E) enhancers at three distances from the promoter and two orientations as indicated in schematics at top of each panel. We also overlay the measurements for both orientations in order to see the influence of orientation (C, F). We use 99 percentile expression in the trunk (0.2-0.9 egg length) to estimate the level of expression driven by each construct. Expression values were normalized by co-staining with endogenous *hkb* (see Materials and Methods) to enable comparison across transgenic lines. Individual embryos are shown as grey dots; bars indicate the mean and 95% confidence interval of the standard error of the mean (SEM). We observe significant differences in expression dependent on distance and orientation. We also thresholded gene expression in the embryos to test whether the position of expression changed. We show an unrolled embryo view for each distance with the percentage of embryos in which a cell expresses the reporter plotted in blue. Cells that were significantly different from the reference line (0bp from promoter in forward orientation) are plotted in red ($p<0.05$, Fisher's Exact Test with permutation to control for multiple hypothesis testing). Position does not change for most lines. The most extreme position shift is a narrowing of the stripes in reverse orientation eve 4/6 at 1000bp from promoter (E).

Expression level varies by as much as 2-fold across the constructs we tested, both in terms of overall level of expression and the relative expression of the two stripes (Figure 2.2 and

Supplemental Figure 2.3). For eve 3/7, expression generally decreases as the enhancer moves away from the promoter, but in the reverse orientation this relationship is not monotonic. Conversely, for eve 4/6, expression increases as the enhancer moves away from the promoter. These results demonstrate that there is a complex relationship between expression level and enhancer position and orientation relative to the promoter.

Despite the changes in expression level, the set of cells expressing the reporter gene was largely consistent in different transgenic lines. After thresholding the gene expression patterns (see Methods), we identified only a handful of cells with statistically significant changes (Figure 2.2). For eve 3/7, these cells are associated with variation in expression along the D/V axis. For eve 4/6, the enhancer drives slightly narrower stripes in the reverse orientation at -1000bp than when it is in the forward orientation adjacent to the promoter. The qualitative similarity of the expression patterns is consistent with previous studies which found that the stripe enhancers drove expression in the appropriate cells even when moved from their endogenous context to a reporter (Small et al. 1992; 1996; Fujioka et al. 1999). These studies used p-element insertions and were therefore limited to qualitative techniques that could accurately measure expression position, but not level. To fully capture the effects of locus organization on gene expression, cellular resolution quantitative methods are required.

Enhancers do not act independently even when separated by a large neutral spacer sequence

We next tested how arrangement and spacing of two enhancers relative to one another and the promoter influences expression pattern. We created a set of constructs using eve 3/7, eve 4/6, and spacer sequences to systematically test the influence of spacing between enhancers (Figure 2.1C). For each spacing we tested several arrangements, labeled A-D, with the spacing indicated by a subscript (Figure 2.1D). Our choice of spacing was based on the distance over which short-range repressors can act, because each *eve* enhancer employs

short-range repressors to direct stripe expression (Clyde et al. 2003; Struffi et al. 2011). Short-range repressors bound at one enhancer are capable of disrupting the activity of another enhancer only if placed within 150bp (Fakhouri et al. 2010). We therefore created constructs where the two enhancers are separated by 1000bp, 200bp, and 0bp.

The eve 3/7 and eve 4/6 enhancers are normally on opposite sides of the gene, separated by approximately 9kb, and thought to act additively (Maeda and Karch 2011). We hypothesized that they would still act additively when both are placed upstream of a reporter gene if separated by a sufficiently large neutral spacer sequence. To test this hypothesis, we created a set of four constructs containing the two enhancers upstream of the promoter with a 1000bp spacer between them, where the orientation and order of the enhancers varies relative to one another and the promoter. Our null expectation was that the output of the two enhancers would simply add together; we calculated this null expectation by adding the expression patterns we measured for each single component enhancer at the properly controlled position and orientation. Comparing the expression patterns driven by our constructs to the null expectation clearly revealed non-additive behavior that depended on the orientation and arrangement of the enhancers (Figure 2.3). The largest discrepancy was for D_{1000}, where expression of stripes 3 and 7 was virtually abolished. In A_{1000}, $B_{1000,}$ and C_{1000} stripe 3 expression increased while stripe 7 did not change, indicating that the two stripes do not always change expression in a coordinated way. The eve 4/6 enhancer had lower than expected expression in A_{1000}, and $B_{1000,}$ but increased slightly in C_{1000}. We conclude that enhancer function is sensitive to the presence of other enhancers in the locus and that the underlying mechanism is affected by the position and orientation of the enhancers relative to one another.

To define the range of influence of enhancers on one another's activity, we moved the enhancers closer to one another in the same four configurations. With a 200bp spacer, we observed additional changes in the level of target gene expression compared to constructs with

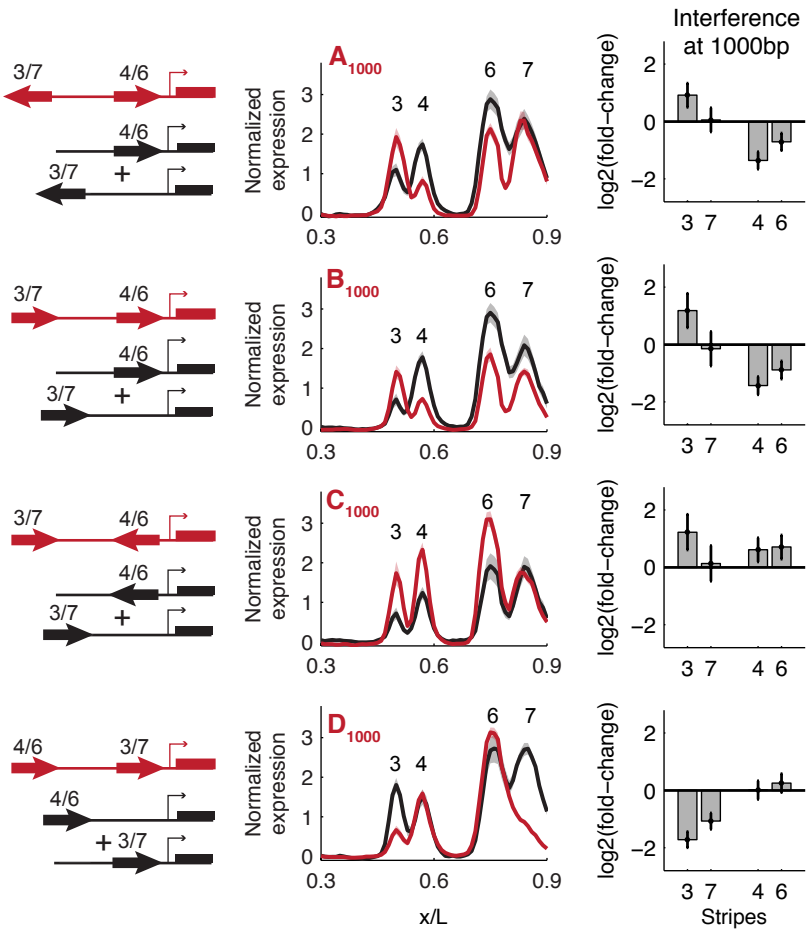


Figure 2.3: Some configurations of enhancers separated by 1000bp produce non-additive expression. We created a set of four constructs containing the two enhancers in different orientations and orders relative to one another with a 1000bp spacer sequence between them. We compared each construct to a null hypothesis of additive activity, as illustrated in the schematics on the left. *hkb* normalized expression as a function of fraction of egg length (x/L) is shown for lateral line traces of test constructs (red), and the null hypothesis (black). Shadows indicate SEM. We also measured the log(fold-change) in mean expression of each stripe relative to the single enhancer controls. Error bars indicate 95% confidence interval of the SEM.

a 1000bp spacer (Figure 2.4). Specifically, the reduced expression of eve 4/6 was even more pronounced in A_{200}, and B_{200}, while D_{200} showed no additional interaction between the two enhancers. In all four configurations, the enhancer closest to the promoter drove lower levels of expression. We conclude that enhancers influence each other's output when they are separated by distances of 200-1000bp, a much longer distance than previously described for interactions between eve 3/7 and eve 2 (Small et al. 1993).

Fused enhancers direct expression patterns only slightly shifted in position

To quantify the influence of short-range interactions between the two enhancers, we fused them together. We expected interactions between the component enhancers to occur at the junctions, due to local interactions between TFs such as short-range repression and

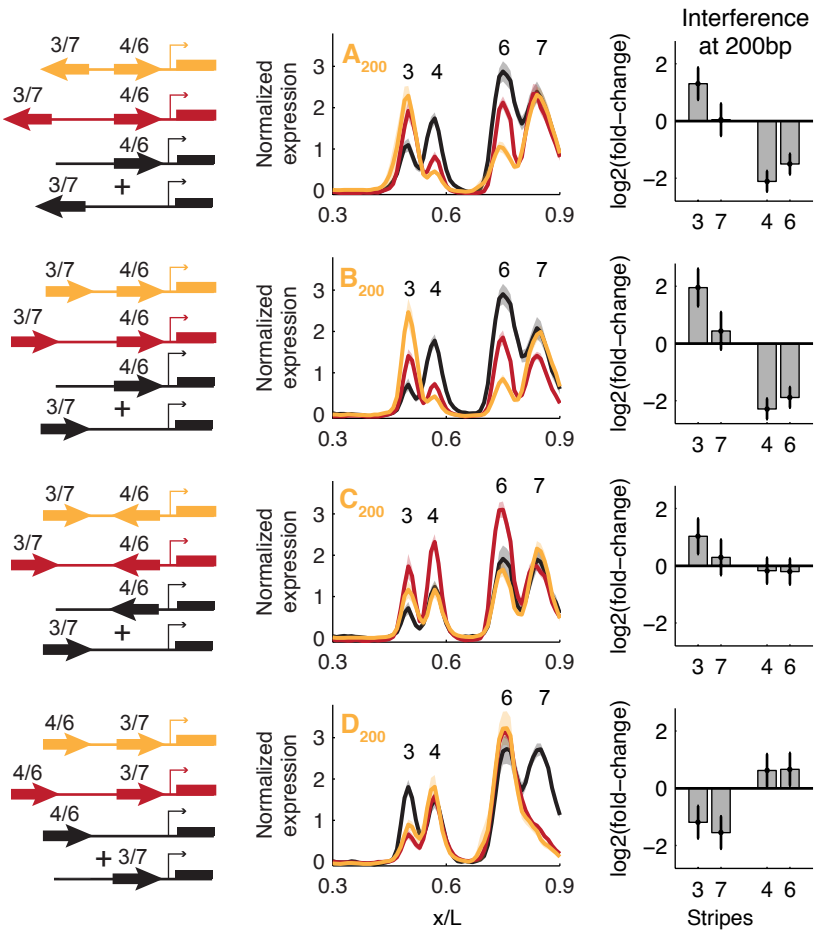


Figure 2.4: Enhancers dramatically influence each other's level when in close proximity. We compared expression from configurations containing a 200bp spacer to 1000bp spacers and the additive null hypothesis. *hkb* normalized expression as a function of fraction of egg length (x/L) is shown for lateral line traces for configurations with a 200bp spacer (yellow), 1000bp spacer (red), and the null hypothesis (black). Shadows indicate SEM. We also plot the log(fold-change) in mean expression of each stripe relative to the single enhancer control. Expression of stripes 4 and 6 are consistently reduced in configurations A, B and C relative to both the 1000bp spacer version and single enhancer controls.

cooperative binding; the four configurations represent all possible junctions between the two enhancers. Previous studies have indicated that short range repression is able to quench activation for up to 150bp on either side of the repressor binding site (Fakhouri et al. 2010; Gray and Levine 1996). Cooperative binding between TFs operates over an even shorter length scale (Crocker et al. 2008; Hanes et al. 1994). Because eve 3/7 and eve 4/6 stripe boundaries are regulated by the same pair of short-range repressors (Clyde et al. 2003; Struffi et al. 2011), we expected that these TFs would act across the junctions of the fused enhancers, thus changing the position of the stripe boundaries driven by this set of reporters (map of TF binding sites in Supplemental Figure 2.1 and Supplemental Figure 2.2).

However, when the eve 3/7 and eve 4/6 enhancers were fused together the position of the stripe boundaries changed only slightly in two configurations; expression level was affected

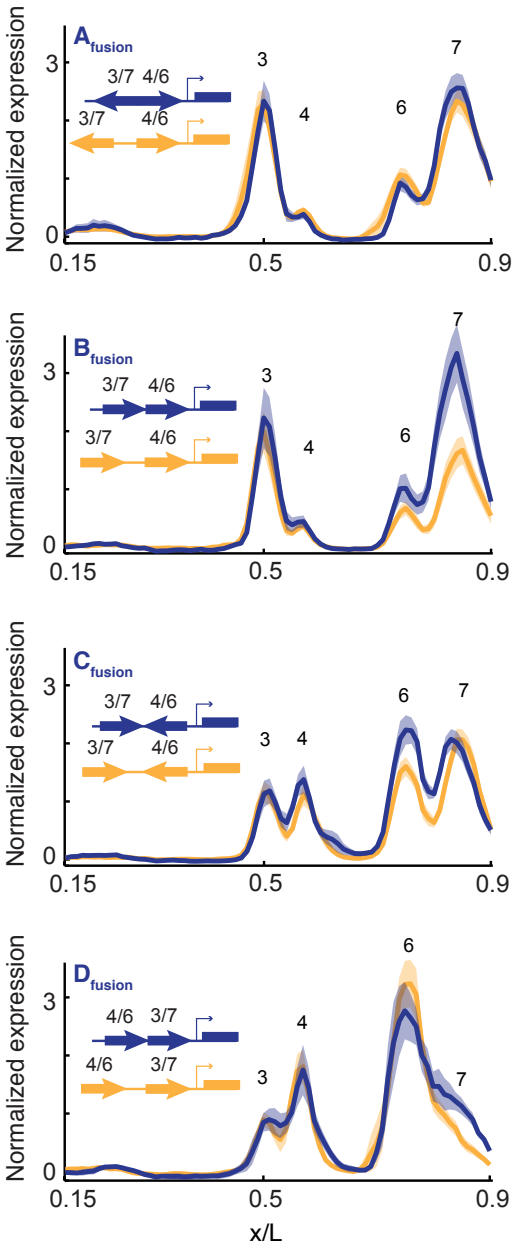


Figure 2.5: Local transcription factor interactions have a minor effect on expression from fused enhancers. We compared fused enhancers (blue) to the same configuration with a 200bp spacer (yellow) to estimate the influence of local interactions between transcription factors bound at the junction on expression. B_{fusion} shows an increase in stripe 7 accompanied by an anterior shift in expression. C_{fusion} shows the same shift in stripe 7 without increased expression. Shadows indicate SEM.

in three configurations (Figure 2.5). We compared the expression driven by enhancers separated by 200bp to those directly juxtaposed in order to compare the influence of locus arrangement to the influence of short-range transcription factor interactions. A_{fusion} exhibited no additional changes in expression level or shifts in expression pattern boundaries. The only expression domain that moved was stripe 7; it shifted anteriorly in C_{fusion} (~1 nucleus width) (Supplemental Figure 2.4). Expression levels of stripes 4 and 6 were slightly higher in B_{fusion} and

C_{fusion}, and expression in the region of stripe 7 was substantially higher in B_{fusion} and slightly increased in D_{fusion}. We conclude that interactions between these two enhancers, even at short distances, predominantly affect expression level, rather than the boundaries of expression patterns.

Levels of gene expression depend on order of enhancers relative to the promoter

Many characteristics of locus architecture can be varied, including order, orientation and spacing of enhancers relative to each other and the promoter. While it is not practical to systematically study all possible architectures even for a simple two enhancer system in intact animals, our data suggest that order of the enhancers relative to the promoter has a significant influence on expression level. The most consistent effect of the enhancer arrangement on target gene expression was a reduction in the level of expression driven by the promoter proximal enhancer. We tested the hypothesis that order relative to the promoter influences the level of expression driven by each enhancer by inverting entire fusions.

Inverting the fusions switched the relative levels of expression driven by the two enhancers. We also observed changes in the relative expression of two stripes driven by the same enhancer (Figure 2.6). $B_{inverted}$ retained high

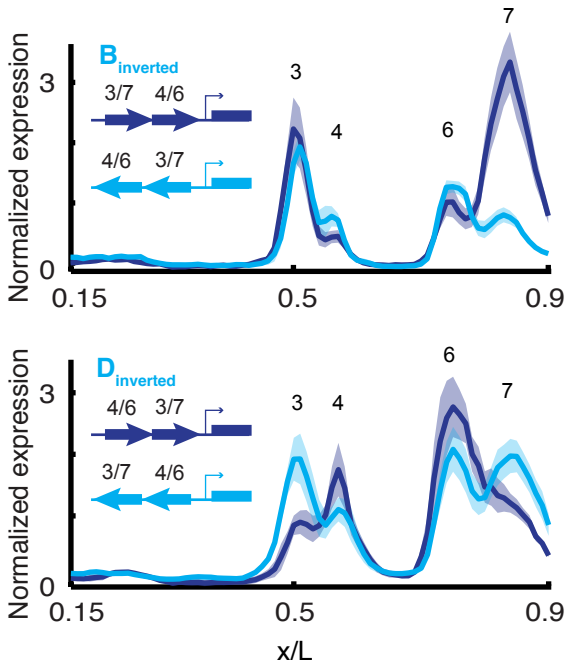


Figure 2.6: Relative proximity to the promoter influences level of expression driven by enhancers. We compared the expression of two fusions (dark blue) to a complete inversion of the entire fusion construct (light blue). In both cases, we see a reversal of the relative levels driven by each component enhancer. Stripes 3 and 7 show large changes in level of expression, while the effect on stripes 4 and 6 is smaller. Shadows indicate SEM.

levels of stripe 3 expression, even as stripe 7 expression was reduced nearly 4-fold. We conclude that order of enhancers has a strong effect on levels of expression, but that other characteristics, such as orientation, also influence level. In combination with our findings from the single enhancer experiments, these results suggest that distance from the promoter needs to be considered both within enhancers, where it manifests as orientation dependence, and across the locus.

Fused enhancers still interact when moved away from the promoter

In all of our constructs one enhancer was immediately adjacent to the promoter. The promoter may exert an influence on enhancer function, either through chromatin or by the basal transcriptional machinery or associated factors. Many promoters have a well positioned nucleosome upstream of the transcription start site (Mavrich et al. 2008), which might occlude portions of the enhancer. Alternatively, the TFs bound at the promoter proximal enhancer could interact directly with the promoter by a different mechanism than when they are farther away. We therefore tested whether moving fused enhancers away from the promoter relieved the repression of the promoter-proximal enhancer.

We found that fusions placed 1000bp upstream of the promoter still drove the same unequal levels of expression as fusions immediately adjacent to the promoter (Figure 2.7). The predominant consequence of moving the fusions away from the promoter was a reduction in the expression in stripes 3 and 7, which is consistent with the observation that the stripe 3/7 enhancer alone had reduced expression when moved away from the promoter. We conclude that depressed expression from the promoter-proximal enhancer does not require a direct juxtaposition of the enhancer and the promoter.

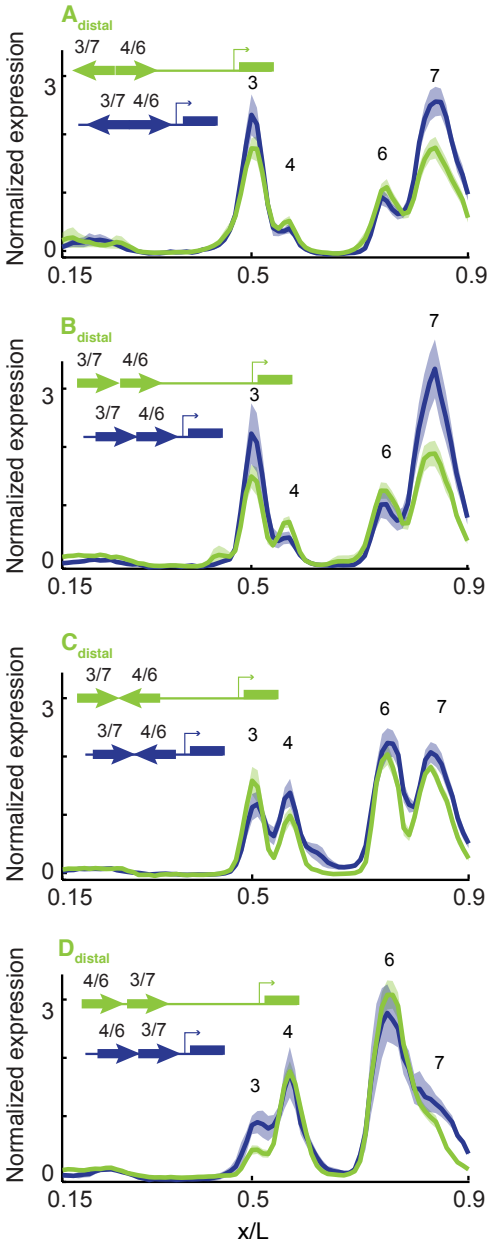


Figure 2.7: Fused enhancers still interact when moved away from the promoter. We tested the role of the promoter environment on determining the level of expression of the proximal enhancer by introducing a 1000bp spacer between fused enhancers and the promoter. Expression of fusions at a distance from the promoter (green) is similar to expression adjacent to the promoter (blue), with the exception of eve 3/7, which is lower in configurations A, B and D, consistent with eve 3/7 having lower expression as it moves away from the promoter. Shadows indicate SEM.

Discussion

To assess the effect of structural variation on gene regulation, we measured reporter gene expression driven by different arrangements of two *eve* enhancers in *Drosophila melanogaster* blastoderm embryos. We systematically varied orientation, order, and spacing of enhancers and measured expression quantitatively at single cell resolution to detect changes in level and position of expression. This approach allowed us to quantify the influence of locus organization while minimizing changes in sequence content. We found that multiple features of locus organization affect expression level significantly; we found only minor changes to the position of the expression pattern. This partially contradicts the classic definition of enhancers as modular units capable of driving the same expression pattern regardless of orientation or distance from the promoter and independent of the activity of other enhancers in the locus (reviewed in Maston et al. 2006). First, we found levels of expression driven by single enhancers vary with the enhancer's orientation and distance from the promoter; the direction and strength of this effect is enhancer-specific. Second, we found that, in constructs containing two enhancers, levels of expression depend on the order and spacing of the enhancers relative to each other. Third, we show that the total expression driven by pairs of enhancers can be non-additive, even when the enhancers are separated by 1000bp, a much longer range expected for short-range repression mechanisms.

Distance between enhancer and promoter influences expression level

Here we show that the expression driven by two single *eve* enhancers is sensitive to the enhancer's position relative to the promoter and this effect is enhancer-specific (Figure 2.2B). These results are consistent with experiments using the SV40 enhancer (Wasylyk et al. 1984) and IFN-beta enhanceosome (Nolis et al. 2009), which found that levels of expression driven by these enhancers also depended on distance from the promoter.

One caveat to this experiment is that the sequence adjacent to the 3' end of the enhancer changes at each distance due to the different lengths of spacer sequence used. It is thus possible that short-range interactions between transcription factors (TFs) bound near the junction between enhancer and spacer sequence are responsible for the distance dependence that we observe. The degeneracy of eukaryotic transcription factor binding motifs (Wunderlich and Mirny 2009) makes it difficult to completely eliminate all TF binding sites in spacer sequences, and the possibility of introducing inappropriate interactions always exists. We examined the predicted TF binding sites in the spacer sequences and found no clear candidates that would explain the observed trends (Supplemental Figure 2.1 and Supplemental Figure 2.2). The spacer sequences contain few predicted TF binding sites, and there are few activator binding sites near the edges of the enhancers, which we would expect to be most strongly affected by short-range interactions. The introduction of long-range repressor binding sites within the spacer could globally decrease expression (Courey and Jia 2001). However, in this case we would expect the effect of distance to be enhancer-independent since the constructs used the same spacers. Instead, the trends are in opposite directions, which suggests that the distance-dependence is not a function of the spacer sequences.

Enhancer-promoter interactions may differ depending on whether an enhancer is promoter-proximal or acting at a distance. Most promoters include a well positioned nucleosome approximately 180bp upstream of the transcription start site (TSS) (Mavrich et al. 2008); when enhancers are in close proximity to the promoter this nucleosome may occlude some binding sites. In yeast, where most regulatory sequences are promoter proximal, nucleosome position has a large effect on which TF binding sites are used (Kim and O'Shea 2008; Raveh-Sadka et al. 2012). In addition, the pre-initiation complex (PIC) containing RNA Pol II, general TFs and co-factors forms a large complex spanning ~100bp across the TSS, and several components have been found to induce DNA bending (reviewed in Levine et al. 2014). Hence, TFs bound to

promoter proximal enhancers can come into direct contact with elements of the PIC (Park and Hong 2012). Conversely, metazoan enhancers commonly act at a distance via looping mediated by mediator, cohesin, and TF binding sites in both the enhancer and promoter (Phillips-Cremins et al. 2013; Kagey et al. 2010; Su et al. 1991). For example, expression of the β-globin gene and looping between the locus control region (LCR) and β-globin promoter are eliminated in GATA1 null cells, but tethering of the two elements with an artificial zinc finger enabled looping and rescued transcription (Deng et al. 2012). The fly *sparkling* enhancer contains a "remote control element" which is required for the enhancer to drive activity at a distance of 846bp, but not when adjacent to the promoter (Swanson et al. 2010). Taken together, these studies support the idea of direct activation when enhancer and promoter are proximal, and a switch to action at a distance mediated by looping.

How might looping result in different levels of expression than direct interaction between enhancers and promoters? It is possible that once looping is established, the interaction between enhancer, bound TFs, and the PIC is the same as when the enhancer is promoter proximal. Thus, the changes in expression level we observe with enhancer-promoter distance may be due to different frequencies or stabilities of enhancer-promoter interactions. However, it is also possible that acting at a distance allows greater conformational freedom and consequently changes the physical interaction of enhancer bound TFs and the promoter in an enhancer-specific manner.

Orientation of enhancer relative to promoter influences expression level

The level of expression driven by individual enhancers in our study is sensitive to enhancer orientation. This is particularly evident when the eve 3/7 enhancer is -500bp from the promoter (Figure 2.2); at this distance the eve 3/7 enhancer drives significantly different levels of expression in each orientation. The relative levels of each stripe driven by a single enhancer

(e.g. stripe 3 and stripe 7) also vary with both distance and orientation (Supplemental Figure 2.1 and Supplemental Figure 2.2). These data demonstrate that the regulatory sequences that generate each stripe are somewhat separable; the stripes need not change in concert. The location of TF binding sites in the enhancer is asymmetric. The orientation dependence may therefore be due to either different TFs coming into contact with TFs bound to the spacer, or the underlying distance-dependence of the TFs interacting with the promoter. Distance-dependent activity for individual binding sites has been demonstrated in both bacteria (Garcia et al. 2012) and yeast (Sharon et al. 2012). Even in these relatively simple systems with single binding sites the distance dependence function is complex. Enhancers contain many TF binding sites, and the aggregate output if each of those binding sites has distance-dependent activity is hard to predict.

In summary, we suggest that the orientation effect is likely due to a combination of asymmetric distribution of binding sites combined with a dependence on distance from the promoter. At minimum, these experiments demonstrate that the information processing in enhancers is asymmetric and highly sensitive to locus context.

Levels of expression depend on the distance and orientation of enhancers relative to each other and the promoter

We found that the largest impact on level of expression was due to interactions between two enhancers in the same reporter construct. The classic definition of enhancers as autonomous units led us to formulate the null hypothesis that the two enhancers would have additive outputs. Contrary to our expectation, we observed a large non-additive interaction effect on level of expression. Our experiments do not address whether the interaction effect is due to direct physical interaction or indirect interaction, for example, through competition for the promoter. However, we can make some observations about the character of the interaction.

The largest effects are correlated with the order of the enhancers relative to the promoter. In general, the enhancer closest to the promoter directs lower expression than expected, while the more distant enhancer directs normal or elevated expression (see Figs 2.3, 2.4 and especially 2.6). The strength of the interaction is dependent on the distance between the enhancers for 3 of 4 cases. The exception to this rule is configuration D_{1000}, in which the repression of eve 3/7 is extremely strong at all distances tested. The magnitude of this effect is much stronger than the effect of short-range interactions between TFs bound at the junctions between fused enhancers (Figure 2.5). Finally, we confirmed that the interaction effect is not due solely to one enhancer being directly adjacent to the promoter (Figure 2.7).

One possible explanation for the observed interaction effect is the formation of direct physical interactions between the enhancers. Many TFs recruit co-factors and adapters for the explicit purpose of establishing long range interactions with the promoter (Phillips-Cremins et al. 2013; Kagey et al. 2010; Su et al. 1991), and these may target other enhancers as well. The two enhancers we used share the same regulating TFs but produce different positions of expression due to different sensitivities to the repressors *hunchback* (*hb*) and *knirps* (*kni*) (Clyde et al. 2003; Struffi et al. 2011). The maintenance of the stripe positions implies that the two enhancers retain separate information integration functions. This constraint argues against the direct interaction of the two enhancers through the formation of a single large complex.

An alternate, indirect, form of interaction between enhancers is through chromatin spreading, which is primarily associated with silencing through long-range repression (Courey and Jia 2001; Li and Arnosti 2011). Some enhancers recruit chromatin-modifying enzymes, which alter the chromatin composition of the locus and might produce either silencing or enhancement of nearby enhancers (discussed in Bulger and Groudine 2011). However, this mechanism would be expected to depend only on presence or absence of a second enhancer, not on the relative arrangement of the two. Even if the chromatin spreading was directional, we

would expect to see an effect that was more strongly dependent on the orientation of the enhancers rather than order relative to the promoter.

An intriguing possibility is that the order of enhancers relative to the promoter may influence the 3D structure of the locus and thus the efficiency of enhancer-promoter interactions. Numerous studies have found correlations between enhancer-promoter looping and gene expression (Deng et al. 2012; Chopra et al. 2012). In addition, a study of the *hox* locus found that enhancers in the locus formed a set of looped contacts even in a transcriptionally silent state, supporting the idea that transcriptionally silent enhancer elements regulate the 3D structure of the locus (Montavon et al. 2011). In our constructs the promoter proximal enhancer may be looped out by the distal enhancer, reducing its expression. However, this explanation does not account for increased expression of the distal enhancer. Most likely we are seeing the combined effects of multiple processes, including regulated looping.

In addition to activating transcription from the promoter, it has recently been shown that enhancers are themselves transcribed (Kim et al. 2010). Enhancer RNAs (eRNA) are generally short-lived, but a variety of putative functional roles have been assigned to them, including recruitment of co-factors and facilitating looping (reviewed in Lam et al. 2014). In yeast, when two promoters drive expression of a single gene, the upstream promoter is used preferentially because transcription through the downstream promoter disrupts its activity (Hirschman et al. 1988; Iyer and Struhl 1995; Martens et al. 2004). It is possible that the interaction between enhancers that we observe is due to a similar effect in which the eRNA produced by one enhancer interferes with the activity of the other.

It is important to note that the two enhancers in our study drive expression in different sets of cells. The existence of an interaction effect therefore indicates that even when enhancers are transcriptionally silent they can influence one another's output. In differentiating cells, enhancers recruit chromatin modifying activity and may interact with basal transcriptional

machinery prior to becoming transcriptionally active (Rada-Iglesias et al. 2011; Creyghton et al. 2010). Our data suggest that "poised" enhancers may influence the activity of neighboring regulatory sequences as well.

Implications for interpreting regulatory sequence variants

Current computational models focus on predicting the activity of single enhancers and do not take locus-level features into account. Single enhancer models are reasonably successful at predicting expression patterns, but do not scale up to the whole locus well (Kim et al. 2013; Samee and Sinha 2014). Using quantitative methods, we have shown that rearrangements of enhancers may affect target gene expression levels, even when binding site content within the enhancer is maintained. In addition, duplications and deletions are likely to have non-additive effects. Our results suggest that including locus-level parameters beyond TF binding will be necessary for accurate predictions.

Implications for regulatory sequence evolution

Given our results that locus organization can affect expression level, selection for expression level may explain conservation of locus architecture. A recent population genetics study in *Drosophila* found that structural variants in both coding and non-coding sequences showed evidence of strong purifying selection (Zichner et al. 2013). Studies in both vertebrates and insects have identified regions of "micro-synteny" in which recombination events are much lower than expected (Sun et al. 2006; Engström et al. 2007; Cande et al. 2009). These regions are enriched for developmental genes and highly conserved elements, a proxy for enhancers. Together, these observations point to an important role for locus architecture in the function of developmental genes.

Materials and Methods

Construction of reporters and transgenic lines

We used RedFly to identify coordinates of the eve stripe 3/7 and stripe 4/6 enhancers (Gallo et al. 2011). The eve_stripe_3+7 element is 510bp (Release 5 coordinates 2R: 5863006-5863516) (Small et al. 1996), while the eve_stripe4_6 element is 800bp (Release 5 coordinates 2R:5871404-5872203) (Fujioka et al. 1999). Note that the stripe 4/6 enhancer coordinates from REDfly contain an extra 208bp on the 3' end compared to the construct tested in Fujioka *et al*. (1999). Enhancers were PCR amplified from genomic DNA from w^{118} *Drosophila melanogaster* flies and sequence verified. Enhancers were inserted into the multiple cloning site of the pBOY vector (Hare et al. 2008) using isothermal assembly (Gibson et al. 2009), which leaves scar-less junctions. LacZ spacer sequences were amplified from the pBOY vector. pBOY contains an *eve* core promoter 20bp downstream of the multiple cloning site that drives an *eve*/ lacZ fusion transcript. The vector also contains an attB site for phiC31 site specific integration (Fish et al. 2007) and the *mini-white* gene for selection of transformants. Each plasmid was injected into attP2 flies (Markstein et al. 2008) by Genetic Services, Inc and transgenic flies were homozygosed using the mini-white eye color marker.

Embryo collection and in situ hybridization

Embryo collection and whole mount *in situ* hybridization was performed as previously described (Luengo Hendriks et al. 2006). Briefly, 0-4hr embryos (25C) were collected, dechorionated in 50% bleach, fixed in a 1:4 mixture of 10% formaldehyde to heptane, and devitellinized in heptane and methanol by shaking. Embryos were post-fixed in formaldehyde and a formaldehyde based hybridization buffer. Hybridizations were performed at 56C with two or three full length cDNA probes: a DIG-labeled probe for *fushi tarazu* (*ftz*), a DNP-labeled lacZ probe and optionally a DNP-labeled probe against *huckebein* (*hkb*). The probes were detected

by successive antibody staining using anti-DIG-HRP (anti-DIG-POD; Roche, Basil, Switzerland) and anti-DNP-HRP (Perkin-Elmer TSA-kit, Waltham, MA, USA), and labeled by reactions with coumarin- and Cy3-tyramide (Perkin-Elmer). Embryos were treated with RNase and incubated with Sytox Green (Invitrogen, Carlsbad, CA, USA) to stain nuclei. Finally, embryos were dehydrated in ethanol and mounted in DePex (Electron Microscopy Sciences, Hatfield, PA, USA), using #1 coverslips to form a bridge to preserve 3D embryo morphology.

Imaging and image processing

Embryos were imaged and computationally segmented for further analysis (Fowlkes et al. 2008). A three-dimensional image stack of each embryo was acquired on a Zeiss LSM Z10 with a plan-apochromat 20x0.8 NA objective using 2-photon microscopy. Embryos were binned into six time points of approximately 10 minute windows using the extent of membrane invagination under phase-microscopy as a morphological marker. Time points correspond to 0-3%, 4-8%, 9-25%, 26-50%, 51-75% and 76-100% membrane invagination along the side of the embryo that has progressed most. Image files were processed into PointCloud representations containing the coordinates and fluorescence levels for each nucleus. Using the *ftz* fiduciary marker, PointClouds were registered to an average morphological template to create a gene expression atlas, a summary text file containing the normalized expression level for each reporter construct in each nucleus at each time point.

hkb normalization

Normalization to a *hkb* co-stain was performed to test the variation in absolute levels of expression across reporters (Wunderlich et al. 2014). Embryos were stained with a mixture of *lacZ*-DNP and *hkb*-DNP probe. Stains were done in two batches: the first batch contained all single enhancer control lines; the second batch contained all two enhancer constructs and two

single enhancer control lines to allow comparison between batches. For each embryo, background was calculated as the mode of the fluorescence distribution. After subtracting background, mean *hkb* fluorescence was calculated as the geometric mean of the anterior and posterior expression domains. We noted that *eve* stripe 7 overlaps slightly with the posterior expression domain of *hkb*, and so chose to use the geometric mean of anterior and posterior rather than solely the posterior domain as in (Wunderlich et al. 2014) to limit the impact of overlapping expression. The fluorescence in each nucleus was then divided by the mean *hkb* fluorescence to yield a normalized expression level.

Data analysis and visualization

Extraction of lateral line traces, and detection of stripe boundaries, and *hkb* normalization were performed in MATLAB using the PointCloud Toolbox (http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp?w=analysis) and custom scripts. Briefly, lateral line traces are a smoothed moving window average over a 1/16th DV strip (about 5 nuclei wide) along the left side of the embryo.

To find predicted TF binding sites shown in Supplemental Figure 2.1 and Supplemental Figure 2.2, we used Patser (http://stormo.wustl.edu/software.html), with PWMs derived from multiple sources (Supplemental Table 2.1). Background GC content was set to 0.406, a P-value limit of 0.001 was used. We plotted the predicted binding sites using InSite, an interactive tool developed by Miriah Meyer (http://www.cs.utah.edu/~miriah/insite/).

Chapter 3: Modeling transcriptional regulation by multiple enhancers

Tara Lydiard-Martin, Md Abul Hassan Samee, Saurabh Sinha, Angela DePace

Author Contributions

TLM, AHD, MAHS and SS designed the computational analysis. MAHS performed the modeling. TLM created figures and analyzed results. TLM and AHD wrote the text.

Introduction

One of the largest remaining frontiers in genome annotation is identification of enhancers (ENCODE Project Consortium et al. 2012; modENCODE Consortium et al. 2010; Kvon et al. 2014; Arnold et al. 2013; Dickel et al. 2014). Because enhancers drive tissue-specific expression during specific time points in development, experimentally screening the entire genome of an organism for enhancer function in all cell types at all time points is a herculean task. Computational models to identify candidate enhancers could greatly facilitate this effort.

One class of such models are called “sequence to expression models”; they predict the expression pattern driven by a specific DNA sequence based on the binding of relevant transcription factors (TFs) (Shea and Ackers 1985; Janssens et al. 2006; Segal et al. 2008; He et al. 2010). These models capture two key steps in gene regulation: a) the probability of TF binding to the DNA sequence based on TF concentration in the cell and measured binding affinity and b) the probability of transcription occurring given a set of bound TFs. Statistical thermodynamics can be used to calculate gene expression from a weighted sum of an ensemble of all possible bound and unbound states (see schematic in Figure 3.1A). This approach has been applied successfully to a variety of gene regulatory sequences from bacteria to metazoans (Bintu et al. 2005; Sherman and Cohen 2012). Extensions of the statistical thermodynamics framework have been used to investigate transcriptional mechanisms in animals including cooperative binding of TFs (He et al. 2010), short-range repression (Fakhouri et al. 2010), nucleosome positioning (Kim and O'Shea 2008), and cooperative co-activation (Kim et al. 2013). To widely employ this type of model, we require only genome sequence and two types of data on relevant TFs—their DNA binding preferences and spatiotemporal expression patterns. Though challenging, it is feasible to attain this type of data for a comprehensive set of TFs in selected tissues (Badis et al. 2009; Segal et al. 2008).

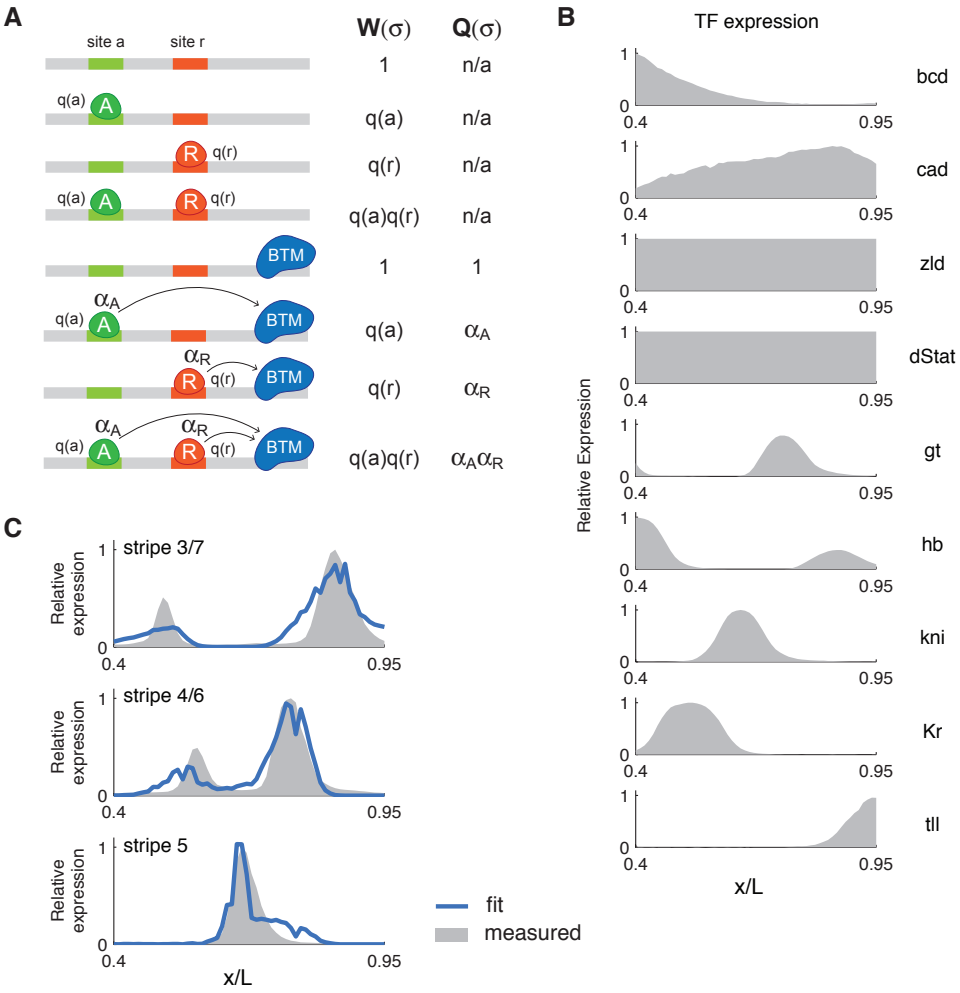


Figure 3.1: Statistical thermodynamic models capture characteristics of expression patterns driven by individual enhancers. A) The GEMSTAT model enumerates all possible states in which TF binding sites (green and red) are occupied or unoccupied and the basal transcriptional machinery (BTM) has assembled. Each state (σ) receives a weight W describing the probability of that state and an activity Q describing the transcriptional activity in that state. The expression in a given cell is proportional to the sum over W*Q for all σ. B) TF expression patterns used in model. C) The direct interaction (DI) model captures domains of stripe expression for individual *eve* enhancers. Measured expression is shown in grey with model fits shown in blue.

To apply sequence to expression models to genome sequence, we must also decide which piece of regulatory DNA to analyze and how to define the boundaries of any piece that is selected as “active.” However, whether enhancers have specific boundaries and how they are formed are still open research questions. Though enhancers are clearly comprised of clusters

of TF binding sites (Berman et al. 2002; Fisher et al. 2012), not all such clusters are active, and TF binding sites are scattered throughout animal genomes (Li et al. 2008; Wunderlich and Mirny 2009). The current leading hypothesis is that interactions between bound TFs define the length of regulatory DNA that will act as a single enhancer (Gray et al. 1994). These interactions occur over short DNA length scales (approximately 150bp) because they require direct protein-protein interactions or binding to a common co-factor (Kulkarni and Arnosti 2005; Fakhouri et al. 2010). Thus, any DNA sequence that contains a gap in TF binding sites longer than 150bp is thought to act as a buffer between clusters of TF binding sites, effectively partitioning them into distinct enhancers.

We tested whether short-range interactions define enhancer boundaries by fusing two developmental enhancers and quantitatively measuring their activity in *Drosophila melanogaster* embryos (Chapter 2). We found that the two enhancers were able to drive distinct spatial expression patterns despite being directly fused; this indicates that the fused enhancers still have some type of "boundary." The two enhancers that we used from the *even-skipped (eve)* locus (eve 3/7 and eve 4/6) both respond to the same repressors *knirps* (*kni*) and *hunchback* (*hb*), but with different sensitivities (Fujioka et al. 1999; Clyde et al. 2003; Struffi et al. 2011). These two enhancers thus interpret the concentration of their regulators differently, and must therefore have some boundary between them. It's thought that the different sensitivities are encoded in the affinity and number of *kni* and *hb* binding sites in each enhancer. Because eve 3/7 and eve 4/6 are on opposite sides of the *eve* locus, a distance far greater than 150bp, short-range repression mechanisms might be sufficient to create a boundary between them in the endogenous context. However, as noted above, they still exhibit a boundary when directly fused, challenging this hypothesis.

Here, we test the hypothesis that short-range repression mechanisms are sufficient to explain enhancer boundaries by fitting three computational model formulations to our dataset of

eve 3/7 and eve 4/6 fusions. Each model is based on a statistical thermodynamics framework initially used to describe bacteriophage lambda (Shea and Ackers 1985) and subsequently extended to describe metazoan enhancers (Janssens et al. 2006; Segal et al. 2008; He et al. 2010). We test all three formulations using the GEMSTAT implementation (He et al. 2010). The first formulation is our null hypothesis—it does not include any short-range interactions between TFs. Instead, in this direct interaction model (DI model), each TF interacts directly with the basal transcriptional machinery to influence the probability of transcription. This model fits a wide range of developmental enhancers with reasonable accuracy (Segal et al. 2008; He et al. 2010). The second formulation is a phenomenological model that tests the idea that drawing explicit enhancer boundaries improves model fits. It uses an iterative approach to identify windows of sequence whose activity contributes to the overall expression profile, and has been successfully applied to identify enhancers in complex developmental gene loci (GL model) (Samee and Sinha 2014). Our third formulation implements a short-range repression mechanism (SR model) and explicitly tests the ability of this mechanism to account for enhancer boundaries in the fusions and constructs containing spacer sequences between the enhancers (He et al. 2010).

We find that although the DI and SR models fit well to individual enhancers, they were unable to capture all four stripes of expression of the fused enhancers. The SR model was able to fit four stripes given sequences with 200bp or 1000bp spacer sequences between the enhancers, consistent with our expectation that the short-range repression mechanism requires buffering sequence between enhancers to provide for enhancer autonomy. The GL model was best able to capture the activity of the fusions, indicating that drawing explicit enhancer boundaries provides the highest quality fit to the data. Notably, GL always selected one window of active sequence within each of the component enhancers, while never overlapping the junction. We propose that an additional mechanism besides short-range repression is involved

in setting the eve 3/7 and eve 4/6 enhancer boundaries, and functions even when enhancers are directly fused.

Results

We fit three different model formulations to a subset of the data described in Chapter 2. To validate the model implementations we first confirmed that each model can simultaneously fit three individual enhancers from the *eve* locus: eve 3/7, eve 4/6 and eve 5. We included the eve 5 enhancer to constrain the parameter values for TFs that have only a small influence on eve 3/7 and eve 4/6. We only considered the trunk of the embryo (0.4-0.95 egg length); this reduces the total number of TFs we need to include and hence the number of parameters to fit. We included TFs that have been specifically shown to regulate the eve 3/7 and eve 4/6 enhancers: *bicoid (bcd), caudal (cad), zelda/vielfeltaig (zld), Stat92e (Dstat), giant (gt), hunchback (hb), knirps (kni), Kruppel (Kr),* and *tailless (tll)*. This set also covers the known regulators of eve 5 (Fujioka et al. 1999).

Next, we challenged each model to fit the expression patterns of four enhancer fusion constructs. To allow the model the greatest chance of finding parameters able to fit the expression pattern, we fit each fusion sequence separately. For the DI and SR models we used a comprehensive method for sampling seed parameter sets and then applied local optimization to the top 2% of seeds (see Methods). GEMSTAT-GL uses a constrained parameter optimization approach (Samee and Sinha 2014). The goodness of fit between experimental data and model prediction was quantified by a score called the “weighted Pattern Generating Potential” (wPGP). This score is different from the original goodness of fit criterion of GEMSTAT (the correlation coefficient), and is described briefly in the Methods and in greater detail in Samee and Sinha (2014). Briefly, the wPGP score more strongly weights cells that are part of the pattern while the correlation coefficient is heavily influenced by small variations in the large number of “off” cells; it is therefore better than the correlation coefficient at fitting narrow patterns such as stripes.

A simple model fits single enhancers but not fusions

The direct interaction (DI) model is the simplest model formulation; it contains only limited interactions between bound TFs. It is similar to non-sequence based models such as linear (Wunderlich et al. 2012) or logistic regressions (Ilsley et al. 2013) with two added constraints. First, the overall regulatory potential of a TF is limited by its capacity to bind the sequence. In non-sequence based models, the regulatory activity of a TF is governed by a parameter that captures both TF strength and occupancy. In the DI model, these two contributions have distinct parameters. Occupancy is predicted by the number and affinity of binding sites in the sequence; TF strength is fit. Second, the DI model does not allow simultaneous occupancy of overlapping TF binding sites. Physical exclusion can explain repression in prokaryotic promoters (Shea and Ackers 1985); such competitive TF binding may also influence overall expression in eukaryotes, though eukaryotic repressors can also function through alternative mechanisms (Li and Arnosti 2011).

In previous work, the DI model was able to fit individual enhancers well (He et al. 2010). First, we tested its ability to simultaneously fit three individual enhancers from the *eve* locus: eve 3/7, eve 4/6 and eve 5. The best fit produced wPGP scores ranging from 0.88 to 0.91 (scores out of a maximum of 1; details in Supplemental Table 3.1). We found that the model was able to capture the domains of stripe expression; it predicted distinct peaks of expression for each stripe, although the predicted expression domain is slightly wider than measured (Figure 3.1C).

We next challenged the DI model using data from the series of enhancer fusions. From sequence characteristics alone, the fusions are indistinguishable from single enhancers—they are 1.3kb in total length and their average TF binding site density is similar to a curated set of 40 *Drosophila* embryonic enhancers (Segal et al. 2008) (Figure 3.2A). We previously observed large variations in the level of expression driven by each component enhancer depending on orientation and order of the enhancers relative to the promoter (Chapter 2). While we did not

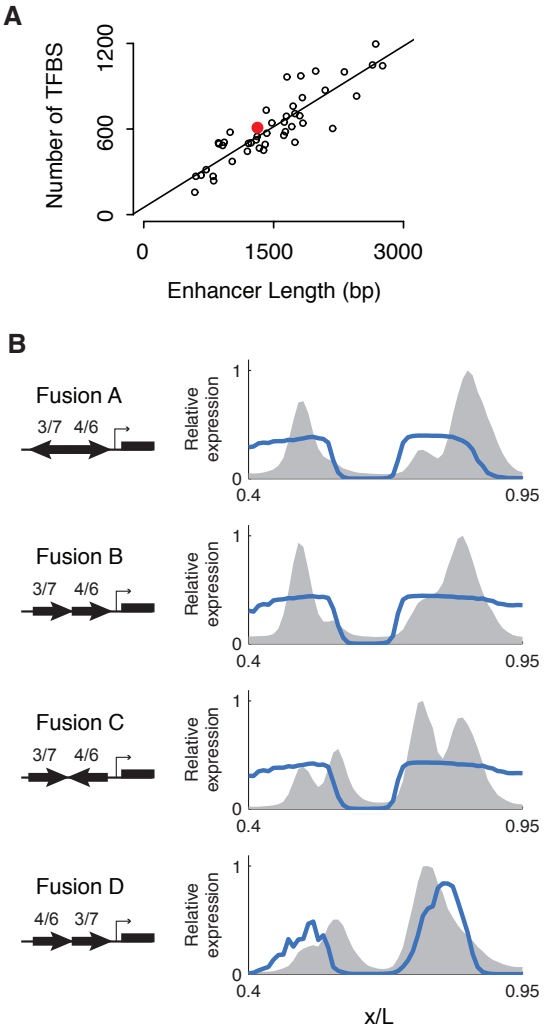


Figure 3.2: Direct Interaction model cannot capture all stripes of fused enhancers. A) Enhancer fusions (red dot) resemble individual enhancers in terms of length and binding site density. We counted predicted binding sites in 40 developmental enhancers (black circles; Segal et al. 2008) and plot number of TFBS against length of the enhancer. Black line shows the fit linear relationship. B) The DI model was fit on each enhancer fusion construct individually to allow maximum model flexibility. Even under the most lenient conditions, the DI model fails to fit four stripes of expression. Measured expression for each fusion is plotted in grey, with fits in blue.

expect the DI model to be able to capture these differences because it does not encode any mechanism for distinguishing enhancer order or orientation, the possibility existed that it might be able to produce four distinct stripes. However, the DI model fails to fit the distinct stripes of fused enhancers. Fit scores ranged from 0.67-0.71 for Fusions A, B and C where the DI model predicts two broad domains of expression. Fusion D had a higher score of 0.80, likely due to the fact that it predicts two more restricted stripes of expression. We attribute the failure of the DI model to averaging across the different sensitivities of each enhancer to the repressors *kni* and *hb.*

Including explicit enhancer boundaries improves model fits

We next fit GEMSTAT-GL, which strongly partitions sequence into active windows that sum to produce the overall expression pattern (Samee and Sinha 2014). This model effectively identifies all individual enhancers in multiple developmental loci, including *eve*, and accurately captures the endogenous gene expression pattern. Predicting the expression pattern of a locus is analogous to predicting the expression pattern driven by the fusions. GEMSTAT-GL must identify the underlying component enhancers, correctly predict their individual expression patterns and how they each contribute to the total expression pattern driven by the fusion. In order to identify active windows of sequence, GEMSTAT-GL uses an iterative approach. It searches for windows whose weighted sum fits the known expression pattern and refits the activity parameters of each TF using the chosen windows. The model scans in 10bp increments for windows ranging in size from 100bp to 1000bp and can select overlapping windows.

GEMSTAT-GL captures the behavior of the enhancer fusions very well, with wPGP scores of 0.97-0.98 (Figure 3.3). Notably, it finds two windows for each fusion, one within each individual enhancer. The windows never overlap or cross the known boundary between enhancers, although this is not a constraint encoded in the model. The individual windows recover the distinct activities of each enhancer, even when the level of expression is so low as to make the stripe hard to distinguish in the total expression pattern as seen in Fusion B (Figure 3.3B). The ability of GEMSTAT-GL to recover distinct expression patterns for each enhancer supports the idea that the enhancers are acting independently *in vivo*.

Short-range repression is not sufficient to capture expression driven by enhancer fusions

Finally, we tested the SR model which encodes the mechanism of short-range repression. Previous work has suggested that short-range repression can encode enhancer boundaries by limiting the range of interactions between TFs to binding sites less than 100bp

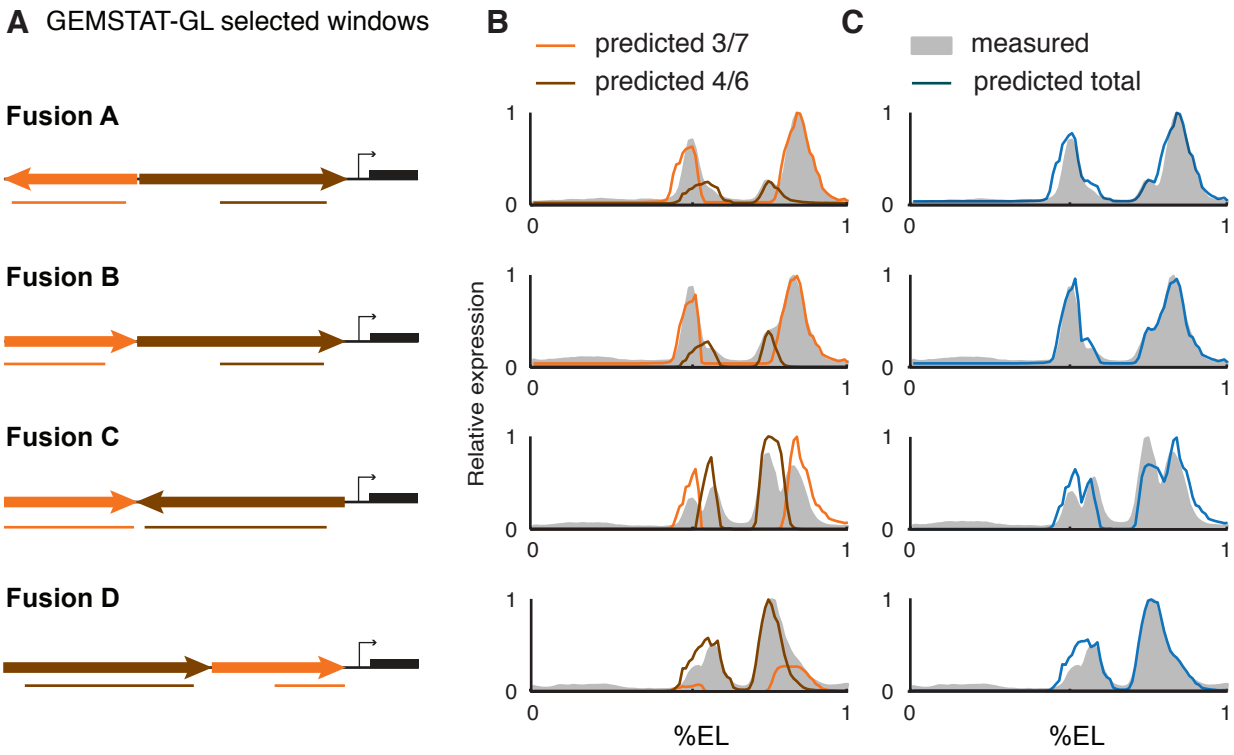


Figure 3.3: GEMSTAT-GL finds windows in each enhancer whose activities sum to produce the observed expression pattern. A) We ran GEMSTAT-GL on a set of four fusion constructs. The discovered windows are indicated by bars under the construct schematic. B) We show the predicted expression for each window in brown and orange. C) Overall fits are shown in blue. In Fusions A, B and D the model excludes the sequence 3' of the junction and includes the complete upstream component enhancer. In Fusion C, the model includes the entirety of the stripe 4/6 enhancer, but as shown in Fusions A and B, the sequence close to the junction is predicted to contribute only a small amount of activity.

apart (Gray et al. 1994). Like the DI model, the SR model also performs well on single enhancers ((He et al. 2010) and Figure 3.4). The single enhancer fits for eve 3/7 and eve 4/6 were slightly improved over the DI model, with scores of 0.94 for both (compared to 0.91 and 0.90 for DI model). In addition, the model was able to capture the four distinct stripes in reporters containing the eve 3/7 and eve 4/6 enhancers separated by either 200bp or 1000bp, as expected (Figure 3.4C). We fit both spacer constructs simultaneously; it is possible that even better fits could be obtained if they were fit separately. The differences in the expression patterns predicted for each construct are due to weak TF binding sites located in the spacer region (Supplemental Figure 3.1).

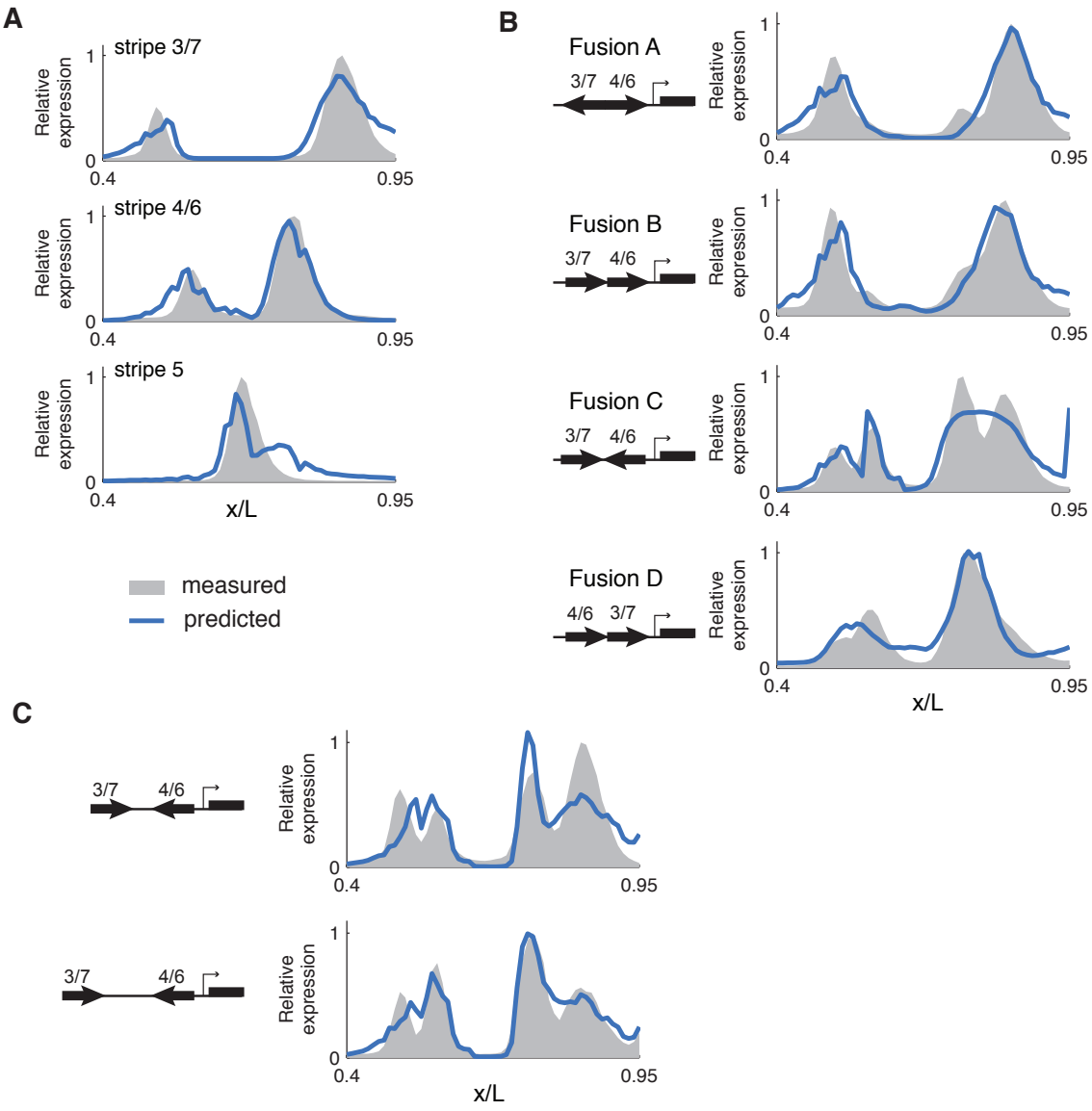


Figure 3.4: SR model cannot capture four stripes in fusions. A) SR model fits stripes of single enhancers. B) SR model only finds two stripes for most fusions, but captures three for Fusion C. It is still not able to distinguish between stripes 6 and 7. C) SR model can form four stripes when a spacer sequence is inserted between enhancers.

In contrast to the good fits for the single enhancers and the enhancers separated by a spacer, we find that the SR model does not fit the fused enhancers as well as the GL model (Figure 3.4B). While the fit scores of 0.90-0.92 are substantially higher than the DI model, the SR model still does not consistently predict 4 stripes, as we observe in our experimental data. Even under the most lenient conditions the model found only two stripes for each fusion except

for Fusion C where it was able to create three stripes. An alternate high-scoring parameter set that fit Fusion C was able to distinguish stripes 6 and 7 but only produced one broad peak covering stripes 3 and 4 (Supplemental Figure 3.2). Rather than fitting the fusions directly, we also tried predicting the expression pattern of the fusion using the parameters fit on individual enhancers or the spacer containing reporters; this also does not produce four stripes (data not shown).

Short-range repression is expected to create enhancer boundaries using spacer sequences between binding sites (Gray et al. 1994). This is consistent with our modeling results in which the SR model is able to fit four stripes of expression from two enhancers when the enhancers are separated by a spacer. However, this mechanism is not sufficient to capture enhancer boundaries when enhancers are fused, suggesting that an additional mechanism is at play.

Discussion

The first sequence to expression models predicted expression of individual enhancers (Janssens et al. 2006; Segal et al. 2008; He et al. 2010). Recent work extends these models to entire loci to address how distinct enhancers each contribute to the overall gene expression pattern of their target. These models allow enhancers to act autonomously, which is a key feature of metazoan gene regulation, either explicitly by defining active windows (Samee and Sinha 2014) or implicitly as a consequence of short-range interactions (Kim et al. 2013). Here we compared the ability of three different sequence to expression models to fit regulatory DNA sequences which contain multiple enhancers with or without spacer sequences between them. This provides a strong test of model performance; all three models could fit the expression patterns of individual enhancers, but only one of them could capture the expression directed by fused enhancers. The most successful model directly partitions regulatory sequence into independent active windows. The model that contains short-range repression is insufficient to fully fit our data, contradicting the prevailing hypothesis that short-range interactions between TFs are sufficient to explain the formation of enhancer boundaries in endogenous loci.

The most successful of our models is phenomenological; it strongly partitions regulatory DNA into independent active sequences without imposing a particular underlying molecular mechanism. Multiple molecular mechanisms could underlie the partitioning; one strong possibility is chromatin state. Including experimental measures of DNA accessibility doubles the accuracy of predicting TF occupancy from sequence and TF occupancy is strongly correlated with active enhancers (Kaplan et al. 2011). Furthermore, some chromatin marks (Calo and Wysocka 2013) and short RNAs (eRNA) (Lam et al. 2014) are correlated with enhancer boundaries, though underlying sequence motifs directing these signals have not yet emerged.

To accurately employ sequence-to-expression models genome-wide, it is not only important to predict which regulatory DNA is active, but also to predict which regulatory DNA will

remain silent even when it contains appropriate TF binding sites. For example, 30% of sequences found to have enhancer activity using a high-throughput screening assay (STARR-seq) were silenced in the endogenous locus (Arnold et al. 2013). This strongly implies that sequence alone is not sufficient to predict activity in the endogenous context. The use of specific windows may be one reason that GL does so well at predicting expression—it is effectively predicting that other regions of the DNA are inaccessible.

The ability of enhancers to encode distinct responses to regulators is critical for patterning. The eve 3/7 and eve 4/6 enhancers used in our study are an example of this phenomenon. Both enhancers have domains of expression limited by *hb* and *kni*, but with different sensitivities so that they form distinct pairs of stripes. We have shown that the prevailing hypothesis that enhancer autonomy is due to short-range repression is insufficient to explain the expression driven by enhancer fusions. In addition, we showed in Chapter 2 that enhancers interact quantitatively even when separated by 1000bp spacer sequences. These higher-order interactions may be due to a variety of mechanisms that will require further investigation to unravel. Understanding these interactions will yield insights into fundamental molecular mechanisms of transcriptional control and improve predictive models to annotate genomes and predict the consequences of regulatory sequence variation.

Materials and Methods

Data collection

For inputs to the model we used previously published TF protein expression data from the FlyEx database (Pisarev et al. 2009). This resource provides high quality protein data for all of the TFs we used in a comparable format to our measurements of reporter mRNA. We obtained the filtered and background subtracted data from cleavage cycle 14A, temporal class 6, which corresponds to our mid-blastoderm stage measurements of reporter expression. Expression of reporter mRNA was measured using fluorescent *in situ* hybridization and 2-photon microscopy as described in Chapter 2.

GEMSTAT

GEMSTAT fits two free parameters for each TF representing 1) the strength of DNA binding and 2) the ability to promote or discourage basal transcriptional machinery (BTM) binding. Finally, a scaling factor is included for each sequence's expression pattern. Although GEMSTAT is able to fit a variety of expression patterns driven by early developmental enhancers (He et al, 2010), the functional form still imposes constraints. Of note for this study, predicted expression changes monotonically with changing TF level; we do not include mechanisms for the strength of TF activity to vary with concentration.

GEMSTAT-GL

GEMSTAT-GL (GEMSTAT for Gene Locus) models a gene's expression as a weighted sum of expression driven by several enhancers within its locus, where each enhancer's output is predicted by GEMSTAT (Samee and Sinha 2014). From the locus sequence, GEMSTAT-GL automatically selects a handful of sequence windows that together generate the gene's expression. It allows the windows to be of varying lengths, even mutually overlapping if

necessary. The number and locations of contributing windows, as well as the weighting factor for each window's contribution are free parameters in the model. During the training phase of the model, we shifted windows by 10bps, and at each starting position window size was varied between 100 bps and 1kb in increments of 10bps. This procedure (a) finds a window whose GEMSTAT readout matches one or more aspects of the gene expression pattern (e.g., stripes), (b) tests if a weighted summation of this window's readout and the readouts of already selected windows improves the overall prediction, and (c) includes the window if such an improvement is noted.

A constrained parameter estimation strategy was used in GEMSTAT-GL to guard against over-fitting. The GEMSTAT model was first trained on ~40 enhancers with A/P patterned expression, while excluding enhancers of the given gene (*eve* in this study). Training on this large data set greatly constrained the model and ruled out over-fitting. The parameter values thus obtained were then used as the starting point of the GEMSTAT-GL parameter training procedure. Thereafter, the training procedure was prohibited from altering any parameter's value by more than two fold from its initial value. This strategy ensured that the final model trained on the given gene was largely consistent with a model that reflects other regulatory parts of the genome. In addition to this conservative approach in fitting the thermodynamic parameters, each pair of weight parameters was also constrained to have a ratio between 1/2 and 2, so that contributions from the different selected windows remained comparable in the final result.

Motif selection and setting LLR threshold

Position weight matrices (PWMs) describing the binding preferences of the regulators are listed in Appendix A Supplemental Table 3.1. We included the activators *bicoid (bcd), caudal (cad), zelda/vielfeltaig (zld), Stat92e (Dstat)*, and repressors *giant (gt), hunchback (hb), knirps*

(kni), Kruppel (Kr), and tailless (tll). A binding site for a regulatory TF was included in the model when its log-likelihood ratio (LLR) score (computed based on the TF's PWM and the background nucleotide distribution) was at least a fraction θ of the LLR score of the TF's strongest binding site. Values of θ were chosen to recover footprinted TF binding sites in the *eve* stripe 2 and stripe 3/7 enhancers taken from RedFly (Gallo et al. 2011).

Parameter fitting of DI and SR models

Given any initialization of parameter values, the GEMSTAT program systematically and iteratively modifies those values and reports a locally optimal parameter setting that maximizes the goodness-of-fit. However, there may exist many other parameter assignments that are as good or nearly as good in terms of their agreement with data, and examining the one optimal assignment reported by GEMSTAT may provide a skewed view of plausible models. We therefore modified the GEMSTAT program to perform a comprehensive exploration of the multi-dimensional parameter space, with the goal of constructing a complete map of plausible quantitative models. To this end, we first generated a large number of N-dimensional vectors (parameter assignments, N = number of parameters) as follows. We partitioned each parameter's allowed range into two halves, which gave us compartments of the parameter space. From each of these compartments, we sampled and scored 1000 vectors of parameter values for their goodness-of-fit to data. We next sorted the 1000 x 2N sampled parameter vectors based on their scores. Finally, for each parameter vector whose score ranks among the top 2% of unique scores in this sorted list we optimized the GEMSTAT model using that vector as initial estimate of model parameters. Starting from an initial set of parameter values, the local optimization routine alternates between the Nelder-Mead simplex method (for 250 iterations) and the quasi-Newton method (for 50 iterations) with w-PGP as the objective function. The chosen parameter sets shown in this Chapter are in Appendix B Supplemental Table 3.2.

wPGP Score

The 'weighted Pattern Generating Potential' (wPGP) score to assess the 'goodness of fit' of a model's predictions (Samee and Sinha 2013) is a modification of the PGP scheme of (Kazemian et al. 2010). Both the schemes were designed to circumvent various shortcomings of the two popular goodness of fit functions, namely the Pearson correlation coefficient and the sum of squared errors. A gene's expression profile consists of the quantitative values of its expression level in different conditions or 'bins'. The essence of wPGP is to compute at each bin 1) a reward for the correctly predicted level of expression and 2) a penalty for over- or under-prediction. The final wPGP score, which ranges between 0 and 1, is a linear combination of these reward and penalty terms across all the bins. Specifically, we let r_i and p_i denote the real and the predicted expression values at bin i. The amount of correctly predicted expression in bin i can then be defined as $\min(p_i, r_i)$, and the reward at bin i is computed as $r_i \times \min(p_i, r_i)$. Thus, the amount of correctly predicted expression is weighted by the real expression level in that bin, and bins with greater expression levels contribute more to the reward term. Similarly, the penalty at any bin is defined as $(1-r_i) \times \text{abs}(p_i - r_i)$. The factor $\text{abs}(p_i - r_i)$, which represents the amount of false prediction (either over- or under-prediction), is thus weighted by the extent of non-expression $(1 - r_i)$ in that bin.

Chapter 4: Discussion

Overturning enhancer modularity

To test assumptions of enhancer modularity, we measured reporter gene expression driven by different arrangements of two *eve* enhancers in *Drosophila melanogaster* blastoderm embryos. We systematically varied orientation, order, and spacing of enhancers and measured expression quantitatively at single cell resolution to detect changes in level and position of expression. This approach allowed us to quantify the influence of locus organization while minimizing changes in sequence content. We found that multiple features of locus organization affect expression level significantly while the position of the expression pattern was largely robust to these changes.

These results partially contradict the classic definition of enhancers as modular units capable of driving the same expression pattern regardless of orientation or distance from the promoter and independent of the activity of other enhancers in the locus (reviewed in Maston et al. 2006). First, we found levels of expression driven by single enhancers vary with the enhancer's orientation and distance from the promoter; the direction and strength of this effect is enhancer-specific. Second, we found that in constructs containing two enhancers, levels of expression depend on the order and spacing of the enhancers relative to each other. We show that the total expression driven by pairs of enhancers can be non-additive, even when the enhancers are separated by 1000bp, a much longer range than expected for known short-range interactions between bound transcription factors. Finally, we assessed the limits of enhancer modularity in direct fusions and found that computational models including known mechanisms of transcription factor interactions cannot account for the observed expression patterns. This indicates that some other mechanism of information processing must be at work in these fusions.

Interpretation of regulatory sequence variants

Our results suggest that structural variants that alter the spacing between enhancers, such as duplications or deletions in intervening sequence, or change the order relative to the promoter, such as inversions, can impact gene expression levels. Such structural sequence variation is quite common in natural populations, and shows signals of negative selection in both protein coding sequence and non-coding regions (Zichner et al. 2013). Both duplications and deletions of enhancers can alter gene expression (Kleinjan and Coutinho 2009), but the impact of other types of variants is less well understood. Our results may provide a functional explanation for regions of micro-synteny in which rearrangements are underrepresented (Engström et al. 2007) and the evolutionary conservation of distances between regulatory elements (Sun et al. 2006; Cande et al. 2009). Alternatively, rearrangements might allow tuning of levels of expression without affecting position of expression.

Flanking sequences may buffer enhancers against changes in locus context. In a BAC containing the entire *eve* locus, deletion of 200bp of sequence flanking the *eve* stripe 2 enhancer caused the gene to become much more sensitive to dosage (i.e. reduced viability of heterozygotes) and extreme temperatures (Ludwig et al. 2011). Although the flanking sequence was not necessary under normal conditions, and did not appear to have direct gene regulatory activity, it clearly contributed to the robustness of expression under genetic and environmental perturbations. In the *sparkling* (*spa*) enhancer, a 5' flanking sequence contains tethering elements necessary for the enhancer to act at a distance (Swanson et al. 2010). Such flanking sequences might also stabilize expression in the case of locus rearrangements. It remains unclear the extent to which other enhancers contain distinct elements mediating enhancer-promoter interactions and whether these elements can buffer locus rearrangements.

A major goal in gene regulation is to be able to predict the consequences of sequence variation on gene expression patterns. Most studies have focused on understanding how sequence variation within an enhancer influences the expression pattern driven by the enhancer. Recent efforts to scale up such models to entire loci have encountered substantial difficulties. Kim et al. fit a mechanistically detailed sequence based model to a set of *eve* stripe 2 and *eve* stripe 3/7 enhancer fusions and used the learned parameters to predict expression across the locus (Kim et al. 2013). While they were able to predict expression of various single enhancers with reasonable accuracy, predicting expression across the locus required hand tuning of a key parameter to produce a reasonable fit. Our collaborators took a different approach by searching for windows of sequence whose predicted output could be summed to produce the total expression pattern of a gene (Samee and Sinha 2014). This approach recovers known enhancers, but requires weights fit to each window to achieve the correct relative levels of expression. Our results suggest that these weights may reflect the effect of locus architecture on gene expression levels. More generally, our results indicate that for sequence-to-expression models to successfully predict output from an intact locus, they must account for genomic context.

Gene regulation in a 3D genome

The 3D shape of the genome influences gene regulation. On a large scale this has been recognized for a long time: active genes co-localize at specific sites at the interior of the nucleus while inactive genes are relegated to the periphery (Fraser and Bickmore 2007) and sites such as polycomb bodies (Pirrotta and Li 2012). Advances in chromatin conformation capture techniques have steadily increased the resolution with which we can view 3D organization of the genome and the regulatory effects of architecture. In megabase scale topologically associating domains (TADs), interactions between sequences are much more frequent than interactions

outside the domain (Dekker et al. 2013). TADs are thought to create boundaries to limit the influence of distal acting regulatory sequences such as enhancers, and genes within a TAD are generally coordinately expressed during development. Although the TAD boundaries themselves are largely invariant during development, interactions within the domain vary extensively as different genes and regulatory sequences become active (Dixon et al. 2012; Sanyal et al. 2012). Evidence is steadily accumulating that such interactions occur prior to the activation of gene expression, suggesting that the establishment of a permissive 3D architecture is a major step in regulation of gene expression (Montavon et al. 2011; Jin et al. 2013; Ghavi-Helm et al. 2014).

Given that regulatory sequences reside in a 3D landscape, our current visualization of enhancers as linear DNA with transcription factor binding sites may be misleading. A better approach may be found by analogy to the annotation and characterization of proteins. Annotation of functional domains along a linear sequence is useful, but structural characterization of domains in 3D provides key insights into function. Protein domains which may not have obvious homology at the sequence level may share structural homology connected to function (Shin et al. 2007). Consideration of the higher order structure of proteins shows that linkers between domains (Wriggers et al. 2005) and interfaces between domains (Panne et al. 2007) can also influence function. If we envision enhancers as regulatory sequence domains, the existence of interactions between domains dependent on ordering and linker length is not surprising, even if we continue to think of enhancers as modules. Thus, in order to understand how linear locus organization influences gene expression, the most effective approach may be to first consider the 3D structure of a gene locus.

Reconstruction of 3D locus architecture is an active area of study (Dekker et al. 2013). A key challenge will be to link the regulatory activity of TFs and structural proteins with the establishment of particular DNA conformations. As with protein folding, this is likely to be a

major challenge. An intermediate goal may be to identify structural classes associated with particular regulatory activities. For example, “super” enhancers consist of tightly clustered enhancers found near key developmental genes (Whyte et al. 2013). These clusters are thought to have particular functional characteristics including heightened sensitivity to certain perturbations. While super enhancers can be identified because they cluster within the linear genome, other enhancers form clusters within 3D space, such as the regulatory archipelago in the *Hox* locus (Montavon et al. 2011). As more data is collected on the 3D architecture of different loci, other structural classes may be identified.

Annotation of regulatory elements

Annotation of regulatory elements is a challenging endeavor. The standard approach is to scan the genome for sequences capable of driving expression in a reporter construct. This can be accomplished in a variety of ways, including validation of computational predictions (Berman et al. 2004), tiling windows (Dickel et al. 2014; Kvon et al. 2014), or recently by using sequencing to detect activity within a library of genomic fragments (STARR-seq) (Arnold et al. 2013). These methods are generally laborious and error prone. In addition, they are limited to identification of sequences which drive expression in the particular cell type or developmental time point assayed. As our results demonstrate, enhancers which do not drive expression in a particular cell type may still influence gene expression. These enhancers will be missed, along with silencers, insulators, tethering sequences and other regulatory elements that modify but do not actively drive expression.

Using functional genomics measurements to annotate sequence may provide a more unbiased approach. Unlike assays that scan for sequences that drive expression in a cell, functional genomics assays such as ChIP-seq and DNase-seq measure binding of proteins to DNA regardless of whether they ultimately drive transcription. A variety of chromatin marks and

co-factors have been associated with regulatory DNA, including promoters, enhancers and "poised" enhancers (Buecker and Wysocka 2012). Poised enhancers do not drive expression in the cell type in which they are identified, but are instead associated with transcriptional activity in earlier or later developmental stages. The same type of poised signature may apply to enhancers expressed in different cells at the same developmental stage (such as eve 3/7 and eve 4/6), but this has not to my knowledge been examined. The ability of functional genomics to annotate these elements is promising. However, association of chromatin marks with regulatory sequences is extremely noisy and correlative. In order to reliably annotate regulatory sequences, the roles of different chromatin modifications and binding of co-factors such as mediator will need to be worked out so that we can distinguish between functional interactions and noise. Predictive sequence to expression models such as those described in Chapter 3 may provide a method for combining different data types to annotate the expected function of a given sequence.

Future directions

We detected interactions between enhancers that direct expression in different cell types by using quantitative techniques to compare both levels and position of expression simultaneously in intact embryos. We isolated the effects of locus organization from those caused by interactions between TF binding sites by creating synthetic arrangements of enhancers and using a neutral spacer sequence taken from the lacZ coding sequence. In this way we improved upon previous work using fusions of the eve 2 and eve 3/7 enhancers which attributed changes in expression patterns to specific TF interactions at the fusion junctions (Small et al. 1993). That work used flanking sequences from the endogenous locus as spacers, which were subsequently shown to contain TF binding sites affecting expression (Kim et al.

2013) as well as a potential role in stabilizing expression under perturbation (Ludwig et al. 2011).

While our approach is powerful, it suffers from several limitations. First, our transgenic reporter constructs are compact; the fusion constructs are 1.3kb in length, with the largest constructs containing a long spacer between the enhancers measuring 2.3kb. Many genomic and biochemical techniques that might help shed insight into the mechanisms involved in enhancer interactions, such as 3C and ChIP, will have poor resolution at these length scales. Secondly, the enhancers we used in the reporter constructs are identical to sequences in the endogenous locus which prohibits using methods such as ChIP that depend on mapping sequencing reads. In the future we could avoid this limitation by using orthologous enhancers from other *Drosophila* species which have conserved function despite extensive sequence evolution (Ludwig et al. 2005; Hare et al. 2008; our unpublished data). Finally, it is not clear to what extent chromatin state and looping might vary across the embryo. If the characteristics are stable, then we could in principle perform measurements in the same system. However, if characteristics of chromatin vary along the embryo, these signals will be averaged in biochemical assays. While these considerations make biochemical assays unappealing at this time, substantial progress is being made to reduce the amount of material required for genomic assays and increase the resolution of measurements (Macaulay and Voet 2014) so that they may be more tractable in the future.

My results highlight the utility of using a synthetic approach to isolate the effects of genomic context on gene expression. It is attractive to consider performing experiments within endogenous loci or BACs with new tools such as CRISPR/Cas9 (Ren:2013dv; Gratz et al. 2014). However, interpreting results from experiments in an endogenous locus is complicated by the unknown role of flanking sequences (as mentioned above) as well as the possibility of regulatory elements such as silencers or insulators within intervening sequences. Many

sequences that do not directly drive expression may have regulatory activity, and it is always possible that an unidentified sequence will impact results in unexpected ways. For these reasons, our lab is pursuing the development of a set of neutral background sequences to use as a synthetic locus in which to perform experiments measuring enhancer interactions. By generating a set of sequences free of known binding sites and assayed for a lack of functional activity, we can ensure that the results of our measurements are interpretable. Using this neutral backbone we can then perform experiments to measure 1) the individual function of each enhancer, 2) how two enhancers interact with endogenous positioning (replace surrounding sequences with a neutral background sequence to avoid confounding interactions with unknown elements in locus), and 3) how multiple enhancers interact with altered spacing and arrangement.

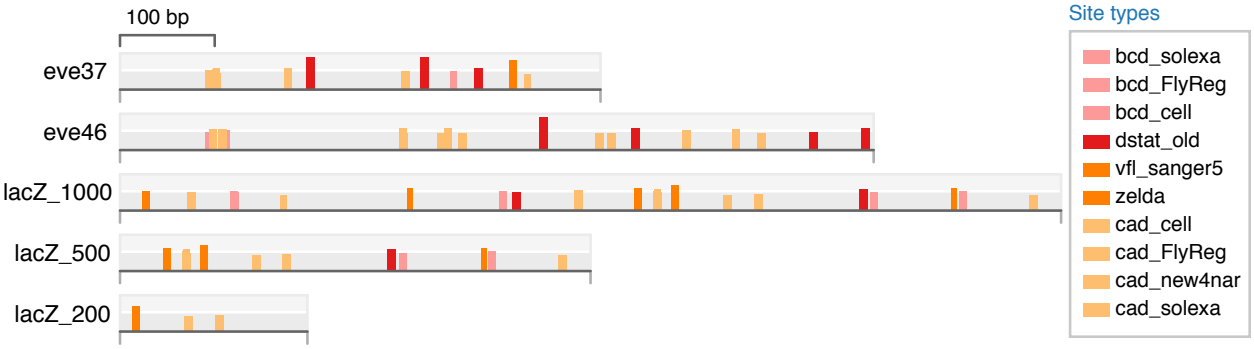
Another critical question raised by my results is whether the TFs responsible for transcriptional activation also govern enhancer-promoter and enhancer-enhancer interactions. Tethering elements have been identified that do not drive expression independently but stabilize enhancer-promoter interactions; similar tethering may be possible between enhancers. We can use our system to screen for factors that influence expression differently in the presence of a second enhancer to dissect the biochemical mechanisms of these interactions.

The fact that locus organization influences level but not position of expression suggests that these two aspects of gene regulation may influence different steps in transcriptional activation. We are adapting the MS2 system, a live reporter of transcriptional activation, to study how interactions between enhancers affect different aspects of transcription, including bursting frequency, rate of initiation, and stochasticity (Garcia et al. 2013). We anticipate that these measurements will allow us to model how different theoretical types of enhancer interactions, including competition and collaboration, influence transcriptional dynamics.

Conclusion

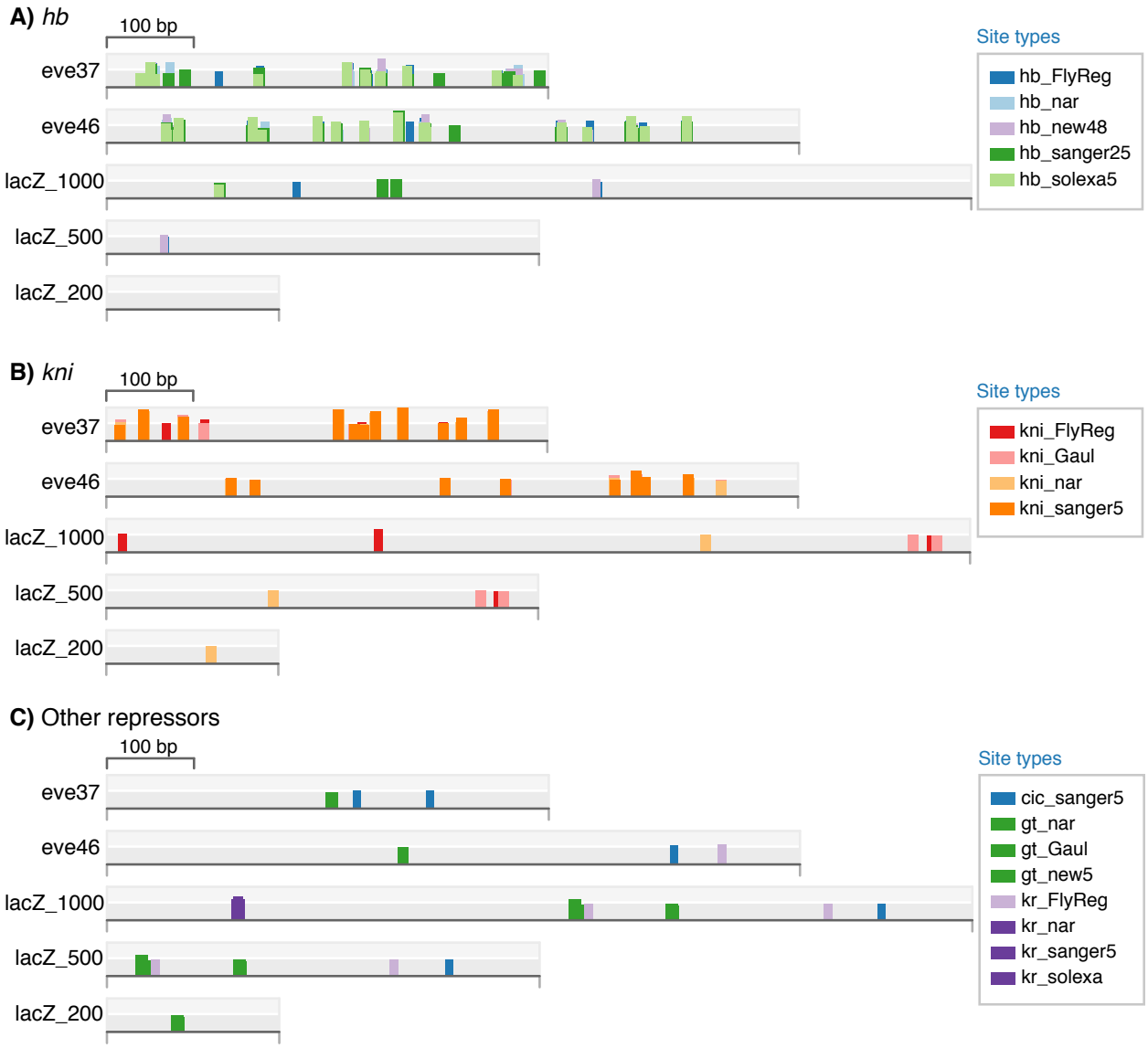
At the outset, I expected to use my early experiments as a basis to explore the mechanisms that produce enhancer modularity. Instead, they revealed a complex role for locus context in enhancer function. Learning the mechanistic basis for this role will require development of new assays to isolate and measure interactions between enhancers and other regulatory sequences. Ultimately, I expect this discovery to lead not only to mechanistic insight, but also refinement of sequence to expression models and the ability to predict gene expression within an endogenous context.

Appendix A: Supplemental Materials for Chapter 2

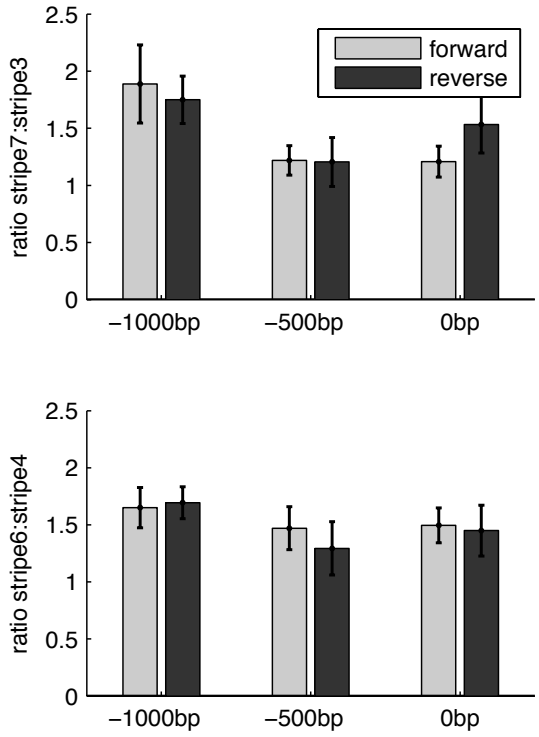


Supplemental Figure 2.1: Activator binding sites in lacZ spacer sequence and enhancers. Predicted activator binding sites found using Patser with a cutoff of p=0.001 are plotted using InSite (see Materials and Methods). The ubiquitous activators *dStat* (red) and *zelda* (orange), as well as *bcd* (pink) and *cad* (light orange). Both enhancers have strong *dStat* sites, and eve 3/7 has one strong *zelda* site, but the activator binding sites are otherwise rather weak. An important note is that eve 3/7 does not have activator binding sites near either enhancer boundary (~90bp on the 5' and ~70bp on the 3' end)--suggesting that short range repressors would need to be very close to the enhancer boundary to influence them. Eve 4/6 also has a large buffer on the 5' end, but none on the 3' end. The 3' end is thought to be ~200bp longer than necessary for full 4/6 expression, although it's possible that those 200bp influence level. The lacZ spacers have quite a few weak activator binding sites. These might be expected to increase background expression, but would not drive spatially localized expression in the absence of repressors to restrict the expression domain.

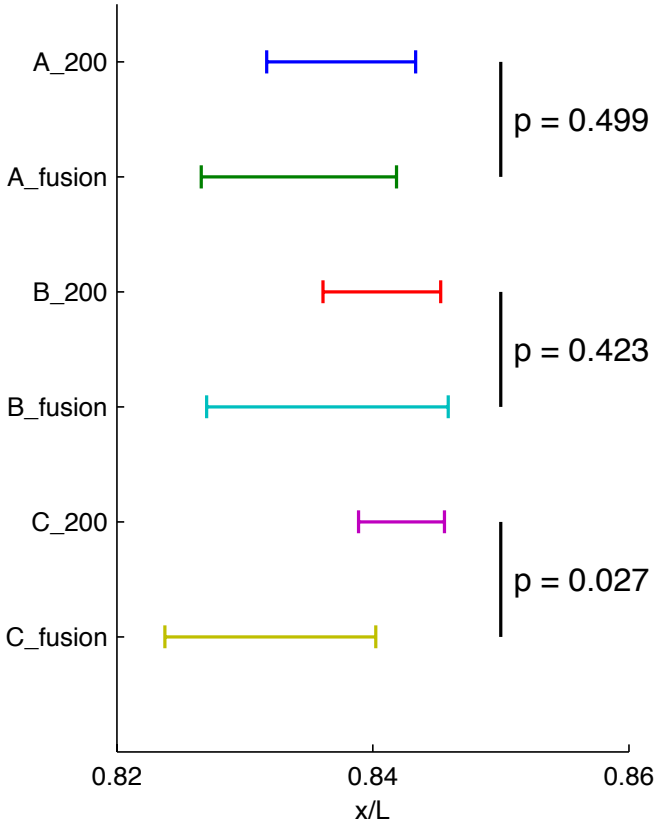
Supplemental Figure 2.2: LacZ spacer sequences are depleted for known repressor binding sites. We first plotted the known 3/7 and 4/6 repressors *hb* (A) and *kni* (B) using multiple PWMs obtained from different types of experimental assays. As expected, eve 3/7 has many more and stronger *kni* binding sites than 4/6. However, it is not obvious that eve 4/6 is more sensitive to *hb* than 3/7. There are a few weak *hb* sites in the 1000bp spacer, but these are not in range to influence the enhancers (the closest to the boundary is ~125bp away). In the lacZ 500bp spacer, there is a single site (only predicted for some of the PWMs) which is close enough to influence the single *dStat* binding site at the 3' end of eve 4/6. The fact that eve 4/6 expression does not change based on orientation in the presence of the 500bp spacer argues against this site influencing expression (Figure 2). The spacers contain more predicted *kni* sites, although they are all of rather low affinity. One cluster is found on the 3' end of both the 1000bp and 500bp spacers. In the single enhancer controls, this cluster would be adjacent to the promoter and may be able to directly repress the promoter. The strongest argument against these sites being active is that the single enhancer controls each behave differently, rather than having consistent repression at the 500bp and 1000bp distances (Figure 2). C) Other repressors: *Capicua* (blue) is a ubiquitous repressor. Other repressors that might influence the expression of the *eve* stripes are *gt* (green) and *Kr* (purple). One of the *Kr* PWMs predicts a number of low affinity sites (light purple), but all three of the other PWMs agree that there is a single moderately strong site in the lacZ 1000bp spacer. That site is 150bp from the boundary of the spacer, and hence unlikely to influence stripe expression. The only *gt* site with the potential to do anything is in lacZ_500 at the 5' end. This is the only repressor site that looks to be in range to potentially influence expression across the boundary.



Supplemental Figure 2.2. LacZ spacer sequences are depleted for known repressor binding sites. (Continued)



Supplemental Figure 2.3: Distance, and to a lesser extent orientation, influence relative levels of expression driven in each stripe by a single enhancer. We measure mean expression in each stripe as in the fold-change plots. Top panel shows stripe 7 to stripe 3 ratios for each of the eve 3/7 constructs. Bottom panel shows stripe 6 to stripe 4 ratios for each of the eve 4/6 constructs. Error bars show 95% confidence interval of the mean.

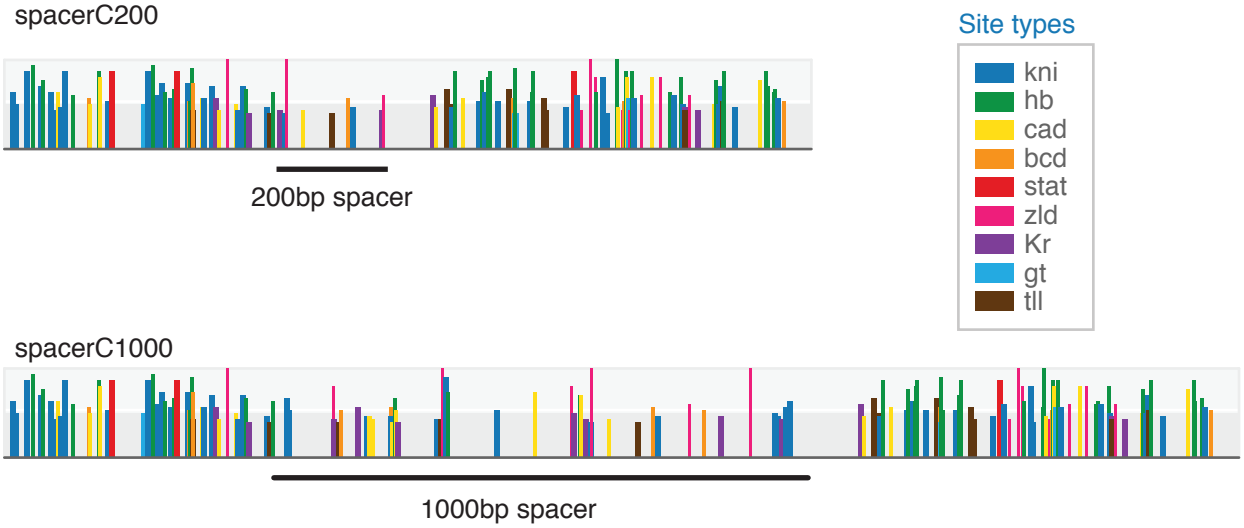


Supplemental Figure 2.4: Stripe 7 peak position shifts slightly in fusions. We measured the AP position of stripe 7 peak expression (x/L) in individual embryos post-registration and plotted the 95% confidence interval of the mean (ie 1.96*SEM). Comparing between the 200bp configurations and fusions, we see an anterior expansion in the range of observed peaks, but only C_fusion is significantly different. The effect is modest and after correcting for multiple hypothesis testing we find that p = 0.08, which fails to meet stringent cutoffs for significance.

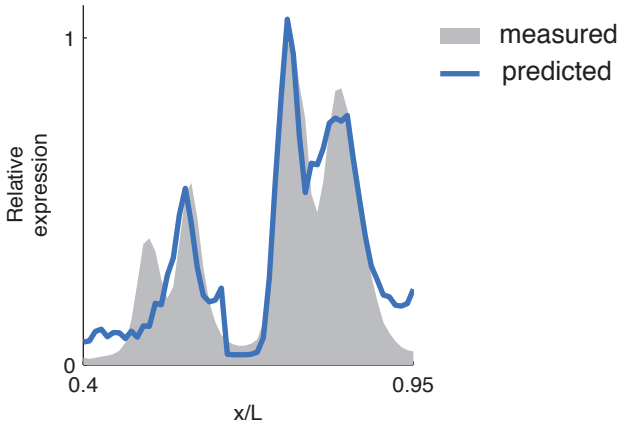
Supplemental Table 2.1: Position weight matrices (PWMs) used in Supplemental Figure 2.1 and Supplemental Figure 2.2 with sources.

TF	PWM Name	Reference
bicoid	bcd_solexa	Noyes et al. 2008a
bicoid	bcd_FlyReg	Bergman et al. 2005
bicoid	bcd_cell	Noyes et al. 2008a
bicoid	bcd_new5NAR	Noyes et al. 2008b
capicua	cic_sanger5	Zhu et al. 2011
caudal	cad_cell	Noyes et al. 2008a
caudal	cad_FlyReg	Bergman et al. 2005
caudal	cad_new4nar	Noyes et al. 2008b
caudal	cad_solexa	Noyes et al. 2008a
dStat	dstat	Noyes et al. 2008b
giant	gt_nar	Noyes et al. 2008b
giant	gt_Gaul	Schroeder et al. 2004
giant	gt_FlyReg	Bergman et al. 2005
giant	gt_new5	Noyes et al. 2008b
hunchback	hb_FlyReg	Bergman et al. 2005
hunchback	hb_nar	Noyes et al. 2008b
hunchback	hb_new48	Noyes et al. 2008b
hunchback	hb_sanger25	Zhu et al. 2011
hunchback	hb_solexa5	Zhu et al. 2011
knirps	kni_flyreg	Bergman et al. 2005
knirps	kni_Gaul	Schroeder et al. 2004
knirps	kni_NAR	Noyes et al. 2008b
knirps	kni_sanger5	Zhu et al. 2011
kruppel	kr_FlyReg	Bergman et al. 2005
kruppel	kr_NAR	Noyes et al. 2008b
kruppel	kr_sanger5	Zhu et al. 2011
kruppel	kr_solexa	Zhu et al. 2011
zelda	vfl_sanger5	Zhu et al. 2011
zelda	zelda	Satija and Bradley 2012

Appendix B: Supplemental Materials for Chapter 3



Supplemental Figure 3.1: TF binding sites in lacZ spacer sequence and enhancers. Predicted TF binding sites found the same LLR cutoff as used in GEMSTAT models (see Materials and Methods). Predicted TF binding sites are illustrated as colored bars where color indicates which TF binds and height represents affinity. The clusters on the ends of the illustrated sequences are the eve3/7 enhancer (left) and eve4/6 enhancer (right). LacZ spacers are depleted for binding sites, but still have some predicted binding.



Supplemental Figure 3.2: An alternate fit of Fusion C was able to produce three stripes of expression. The fit shown captures distinct stripe 6 and 7 expression patterns, but excludes stripe 3. Both this fit and the one shown in main text had wPGP scores of 0.91. Measured expression is shown in grey with fit shown in blue.

Supplemental Table 3.1: wPGP Scores for model fits

DI	wPGP
eve37	0.914979
eve46	0.905901
eve5	0.876743
fusionA	0.673055
fusionB	0.711058
fusionC	0.679346
fusionD	0.797772
spacerC1000	0.813482
spacerC200	0.753688
SRR	
eve37	0.942237
eve46	0.937873
eve5	0.85518
fusionA	0.925779
fusionB	0.920015
fusionC_1031	0.908841
fusionC_1045	0.906202
fusionD	0.919115
spacerC_1000_1768	0.92919
spacerC_200_1768	0.821166
GEMSTAT-GL	
Fusion A	0.97424
Fusion B	0.96758
Fusion C	0.97149
Fusion D	0.976795
SpacerC200	0.96829
SpacerC1000	0.94971

Supplemental Table 3.2: Parameter values used in models

		single	fusionA	fusionB	fusionC1	fusionC2	fusionD	spacer1000	spacer200
max DNA binding	bcd	0.12	1.73	26.70	0.01	0.01	0.01	0.01	0.01
	cad	5.46	0.26	0.32	4.45	25.54	1.33	3.27	3.27
	zld	7.50	0.06	0.32	0.86	0.22	0.61	4.33	4.33
	stat	0.95	5028.20	0.04	0.01	5868.39	0.02	0.44	0.44
	gt	0.88	10.72	0.02	5000.31	4826.12	1.10	5113.65	5113.65
	hb	11.53	0.79	0.02	0.18	114.60	0.01	1.00	1.00
	kni	0.08	0.02	3.82	0.17	0.05	0.06	13.06	13.06
	Kr	1.41	0.01	0.04	0.83	0.26	0.10	0.01	0.01
	tll	1.28	0.69	0.02	0.09	2.81	0.37	1.47	1.47
txp effect	bcd	1.00	10.25	10.25	1.00	9.69	1.00	1.00	1.00
	cad	6.58	10.25	10.25	10.25	1.09	1.00	9.53	9.53
	zld	1.00	10.25	10.10	3.24	10.25	10.25	1.45	1.45
	stat	1.00	1.00	9.72	1.00	1.33	1.00	10.25	10.25
	gt	3387.04	3576.17	99573.09	10000.00	170.02	453.78	179.38	179.38
	hb	1.00	1.00	54.49	29.66	1.44	404.00	134.21	134.21
	kni	425.52	2026.78	4.66	739.60	561.02	128.46	11.22	11.22
	Kr	120.47	1.00	14996.93	743.82	31831.11	100000.0	265.14	265.14
	tll	5602.56	1285.51	54004.56	99007.56	687.11	1518.36	1022.23	1022.23
	basal	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
coop	bcd-bcd	88.72	0.01	0.01	100.50	100.41	90.13	0.01	0.01
	cad-cad	3.78	100.34	98.28	100.50	0.01	100.50	7.90	7.90

Bibliography

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**: 685–692.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077.

Arnosti DN, Barolo S, Levine M, Small S. 1996. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**: 205–214.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**: 1720–1723.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bergman CM, Carlson JW, Celniker SE. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* **21**: 1747–1749.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci* **99**: 757–762.

Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE. 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**: R61.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. 2005. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev* **15**: 125–135.

Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**: 276–284.

Bulger M, Groudine M. 2010. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol* **339**: 250–257.

Bulger M, Groudine M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**: 327–339.

Calo E, Wysocka J. 2013. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**: 825–837.

Cande J, Goltsev Y, Levine MS. 2009. Conservation of enhancer location in divergent insects. *Proc Natl Acad Sci* **106**: 14414–14419.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**: 577–580.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**: 302–305.

Chopra VS, Kong N, Levine M. 2012. Transcriptional repression via antilooping in the Drosophila embryo. *Proc Natl Acad Sci* **109**: 9460–9464.

Clyde DE, Corado MSG, Wu X, Pare A, Papatsenko D, Small S. 2003. A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature* **426**: 849–853.

Courey AJ, Jia S. 2001. Transcriptional repression: the long and the short of it. *Genes Dev* **15**: 2786–2796.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* **107**: 21931–21936.

Crocker J, Tamori Y, Erives A. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6**: e263.

de Laat W, Duboule D. 2013. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**: 499–506.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**: 390–403.

Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA. 2012. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**: 1233–1244.

Dickel DE, Zhu Y, Nord AS, Wylie JN, Akiyama JA, Afzal V, Plajzer-Frick I, Kirkpatrick A, Göttgens B, Bruneau BG, et al. 2014. Function-based identification of mammalian enhancers using site-specific integration. *Nat Meth* **11**: 566–571.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Dunipace L, Ozdemir A, Stathopoulos A. 2011. Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. *Development* **138**: 4075–4084.

ENCODE Project Consortium, Dunham I, Khatun J, Kundaje A, Birney E, Green ED, Bernstein BE, Gerstein M, Hardison RC, Snyder M, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res* **17**: 1898–1908.

Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EEM. 2014. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. ed. M. Levine. *PLoS Genet* **10**: e1004060.

Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Arnosti DN. 2010. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol Syst Biol* **6**: 341.

Fish MP, Groth AC, Calos MP, Nusse R. 2007. Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. *Nat Prot* **2**: 2325–2331.

Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, et al. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci* **109**: 21330–21335.

Fowlkes CC, Hendriks CLL, Keränen SVE, Weber GH, Rübel O, Huang M-Y, Chatoor S, DePace AH, Simirenko L, Henriquez C, et al. 2008. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell* **133**: 364–374.

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* **474**: 598–603.

Fraser P, Bickmore W. 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**: 413–417.

Fujioka M. 2010. Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors. *Dev Biol* **344**: 1048–1059.

Fujioka M, Emi-Sarker Y, Yusibova GL, Goto T, Jaynes JB. 1999. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **126**: 2527–2538.

Gallo SM, Gerrard DT, Miner D, Simich M, Soye Des B, Bergman CM, Halfon MS. 2011. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res* **39**: D118–23.

Garcia HG, Sanchez A, Boedicker JQ, Osborne M, Gelles J, Kondev J, Phillips R. 2012. Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Rep* **2**: 150–161.

Garcia HG, Tikhonov M, Lin A, Gregor T. 2013. Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning. *Curr Biol* **23**: 2140–2145.

Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EEM. 2014. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*. NA

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth* **6**: 343–345.

Gillies SD, Morrison SL, Oi VT, Tonegawa S. 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**: 717–728.

Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, O'Connor-Giles KM. 2014. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. *Genetics* **196**: 961–971.

Gray S, Levine M. 1996. Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. *Genes Dev* **10**: 700–710.

Gray S, Szymanski P, Levine M. 1994. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev* **8**: 1829–1838.

Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics* **166**: 1775–1782.

Guenther C, Pantalena-Filho L, Kingsley DM. 2008. Shaping skeletal growth by modular regulatory elements in the Bmp5 gene. ed. D.R. Beier. *PLoS Genet* **4**: e1000308.

Hanes SD, Riddihough G, Ish-Horowicz D, Brent R. 1994. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. *Mol Cell Biol* **14**: 3364–3375.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. ed. N. Perrimon. *PLoS Genet* **4**: e1000106.

He X, Samee MAH, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. ed. U. Ohler. *PLoS Comput Biol* **6**: e1000935.

Hirschman JE, Durbin KJ, Winston F. 1988. Genetic evidence for promoter competition in Saccharomyces cerevisiae. *Mol Cell Biol* **8**: 4608–4615.

Ilsley GR, Fisher J, Apweiler R, DePace AH, Luscombe NM. 2013. Cellular resolution models for even skipped regulation in the entire Drosophila embryo. *Elife* **2**: e00522–e00522.

Iyer V, Struhl K. 1995. Mechanism of differential utilization of the his3 TR and TC TATA elements. *Mol Cell Biol* **15**: 7059–7066.

Janssens H, Hou S, Jaeger J, Kim A-R, Myasnikova E, Sharp D, Reinitz J. 2006. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nat Genet* **38**: 1159–1165.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–294.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430–435.

Kaplan T, Li X-Y, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, Eisen MB. 2011. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development ed. G.S. Barsh. *PLoS Genet* **7**: e1001290.

Karczewski KJ, Dudley JT, Kukurba KR, Chen R, Butte AJ, Montgomery SB, Snyder M. 2013. Systematic functional regulatory assessment of disease-associated variants. *Proc Natl Acad Sci* **110**: 9607–9612.

Kazemian M, Blatti C, Richards A, McCutchan M, Wakabayashi-Ito N, Hammonds AS, Celniker SE, Kumar S, Wolfe SA, Brodsky MH, et al. 2010. Quantitative Analysis of the Drosophila Segmentation Regulatory Network Using Pattern Generating Potentials. *PLoS Biol* **8**: e1000456.

Kim A-R, Martinez C, Ionides J, Ramos AF, Ludwig MZ, Ogawa N, Sharp DH, Reinitz J. 2013. Rearrangements of 2.5 kilobases of noncoding DNA from the Drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. ed. M. Levine. *PLoS Genet* **9**: e1003243.

Kim HD, O'Shea EK. 2008. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* **15**: 1192–1198.

Kim MJ, Oksenberg N, Hoffmann TJ, Vaisse C, Ahituv N. 2014. Functional characterization of SIM1-associated enhancers. *Hum Mol Genet* **23**: 1700–1708.

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187.

Kirschner M, Gerhart J. 1998. Evolvability. *Proc Natl Acad Sci USA* **95**: 8420–8427.

Kleinjan D-J, Coutinho P. 2009. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. *Brief Funct Genomic Proteomic* **8**: 317–332.

Klopocki E, Ott C-E, Benatar N, Ullmann R, Mundlos S, Lehmann K. 2008. A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. *J Med Genet* **45**: 370–375.

Kulkarni MM, Arnosti DN. 2005. cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. *Mol Cell Biol* **25**: 3411–3420.

Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A. 2014. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci* **109**: 19498–19503.

Lam MTY, Li W, Rosenfeld MG, Glass CK. 2014. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39**: 170–182.

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**: R754–63.

Levine M, Cattoglio C, Tjian R. 2014. Looping Back to Leap Forward: Transcription Enters a New Era. *Cell* **157**: 13–25.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84–98.

Li LM, Arnosti DN. 2011. Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. *Curr Biol* **21**: 406–412.

Li X-Y, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**: e27.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**: 564–567.

Ludwig MZ, Manu, Kittler R, White KP, Kreitman M. 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. ed. H.S. Malik. *PLoS Genet* **7**: e1002364.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol* **3**: e93.

Luengo Hendriks CL, Keränen SVE, Fowlkes CC, Simirenko L, Weber GH, DePace AH, Henriquez C, Kaszuba DW, Hamann B, Eisen MB, et al. 2006. Three-dimensional

morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol* **7**: R123.

Macaulay IC, Voet T. 2014. Single cell genomics: advances and future perspectives. *PLoS Genet* **10**: e1004126.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Maeda RK, Karch F. 2011. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Curr Opin Genet Dev* **21**: 187–193.

Mallarino R, Grant PR, Grant BR, Herrel A, Kuo WP, Abzhanov A. 2011. Two developmental modules establish 3D beak-shape variation in Darwin's finches. *Proc Natl Acad Sci* **108**: 4057–4062.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE. 2011. The developmental role of Agouti in color pattern evolution. *Science* **331**: 1062–1065.

Markstein M, Pitsouli C, Villalta C, Celniker SE, Perrimon N. 2008. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat Genet* **40**: 476–483.

Martens JA, Laprade L, Winston F. 2004. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429**: 571–574.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29–59.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**: 1190–1195.

Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, et al. 2008. Nucleosome organization in the Drosophila genome. *Nature* **453**: 358–362.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**: 271–277.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Nègre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, de Laat W, Spitz F, Duboule D. 2011. A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**: 1132–1145.

Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**: 7058–7076.

Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D. 2009. Transcription factors mediate long-range enhancer-promoter interactions. *Proc Natl Acad Sci* **106**: 20222–20227.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008a. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**: 1277–1289.

Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008b. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* **36**: 2547–2560.

Nüsslein-Volhard C, Wieschaus E. 1980. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**: 795–801.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**: 283–293.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**: R52.

Panne D, Maniatis T, Harrison SC. 2007. An atomic model of the interferon-beta enhanceosome. *Cell* **129**: 1111–1123.

Park KW, Hong J-W. 2012. Mesodermal repression of single-minded in Drosophila embryo is mediated by a cluster of Snail-binding sites proximal to the early promoter. *BMB Rep* **45**: 577–582.

Perry MW, Boettiger AN, Bothma JP, Levine M. 2010. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr Biol* **20**: 1562–1567.

Perry MW, Boettiger AN, Levine M. 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci* **108**: 13570–13575.

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**: 1281–1295.

Pirrotta V, Li H-B. 2012. A view of nuclear Polycomb bodies. *Curr Opin Genet Dev* **22**: 101–109.

Pisarev A, Poustelnikova E, Samsonova M, Reinitz J. 2009. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Res* **37**: D560–6.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**: 279–283.

Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* **44**: 743–750.

Romano LA, Wray GA. 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* **130**: 4187–4199.

Samee AH, Sinha S. 2013. Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data. *Methods* **62**: 79–90.

Samee MAH, Sinha S. 2014. Quantitative modeling of a gene's expression from its intergenic sequence. ed. A. Tanay. *PLoS Comput Biol* **10**: e1003467.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**: 109–113.

Satija R, Bradley RK. 2012. The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the Drosophila embryo. *Genome Res* **22**: 656–665.

Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. 2004. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2**: E271.

Segal E, Raveh-Sadka T, Schroeder MD, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**: 535–540.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology* **30**: 521–530.

Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181**: 211–230.

Sherman MS, Cohen BA. 2012. Thermodynamic state ensemble models of cis-regulation. *PLoS Comput Biol* **8**: e1002407.

Shin DH, Hou J, Chandonia J-M, Das D, Choi I-G, Kim R, Kim S-H. 2007. Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center. *J Struct Funct Genomics* **8**: 99–105.

Small S, Arnosti DN, Levine M. 1993. Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development* **119**: 762–772.

Small S, Blair A, Levine M. 1992. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* **11**: 4047–4057.

Small S, Blair A, Levine M. 1996. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Dev Biol* **175**: 314–324.

Small S, Kraut R, Hoey T, Warrior R, Levine MS. 1991. Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes Dev* **5**: 827–839.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**: 1021–1028.

Struffi P, Corado M, Kaplan L, Yu D, Rushlow C, Small S. 2011. Combinatorial activation and concentration-dependent repression of the Drosophila even skipped stripe 3+7 enhancer. *Development* **138**: 4291–4299.

Su W, Jackson S, Tjian R, Echols H. 1991. DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev* **5**: 820–826.

Sun H, Skogerbø G, Chen R. 2006. Conserved distances between vertebrate highly conserved elements. *Hum Mol Genet* **15**: 2911–2922.

Swanson CI, Evans NC, Barolo S. 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* **18**: 359–370.

Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr Biol* **21**: 1186–1196.

Thanos D, Maniatis T. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**: 1091–1100.

VanderMeer JE, Ahituv N. 2011. cis-regulatory mutations are a genetic cause of human limb malformations. eds. M.A. Ros and J.F. Fallon. *Dev Dyn* **240**: 920–930.

Venken KJT, Carlson JW, Schulze KL, Pan H, He Y, Spokony R, Wan KH, Koriabine M, de Jong PJ, White KP, et al. 2009. Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. *Nat Meth* **6**: 431–434.

Wasylyk B, Wasylyk C, Chambon P. 1984. Short and long range activation by the SV40 enhancer. *Nucleic Acids Res* **12**: 5589–5608.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci* **110**: 11952–11957.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**: 307–319.

Wittkopp PJ, Stewart EE, Arnold LL, Neidert AH, Haerum BK, Thompson EM, Akhras S, Smith-Winberry G, Shefner L. 2009. Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. *Science* **326**: 540–544.

Wriggers W, Chakravarty S, Jennings PA. 2005. Control of protein functional dynamics by peptide linkers. *Biopolymers* **80**: 736–746.

Wunderlich Z, Bragdon MD, DePace AH. 2014. Comparing mRNA levels using in situ hybridization of a target gene and co-stain. *Methods* **68**: 233–241.

Wunderlich Z, Bragdon MD, Eckenrode KB, Lydiard-Martin T, Pearl-Waserman S, DePace AH. 2012. Dissecting sources of quantitative gene expression pattern divergence between Drosophila species. *Mol Syst Biol* **8**: 604.

Wunderlich Z, Mirny LA. 2009. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* **25**: 434–440.

Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, et al. 2011. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res* **39**: D111–7.

Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO. 2013. Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Res* **23**: 568–579.